THIS WEEK

Power to the people

A planned database collating medical information for England's population is a laudable exercise, with huge potential for research. But people's right to opt out has been greatly downplayed.

This month, the National Health Service (NHS) in England will bombard all 26.5 million households in the country with an innocuous-looking information flyer. In soothing tones the leaflet, entitled 'Better Information Means Better Care', announces changes to the way that health officials will handle confidential medical records to "improve the quality of care and health services for all".

The name of the programme is care.data — not that you would know that from the leaflet, which neglects to mention it. From this spring, information from medical consultations, for example on diagnoses and treatments prescribed, that was once confidential between a person and their doctor, will be uploaded to a central database. There it will be combined with hospital and other medical records. This will become one of the world's most complete databases on the health care of patients. Initially, the data will be used to help health authorities to manage NHS resources, but the plan is to eventually open up the database to researchers and private companies. The importance for research and the medical opportunities afforded by such a unique joined-up resource cannot be overestimated. *Nature* fully supports such an endeavour, and this journal spoke out in May 2013 when proposed changes to European data-protection laws threatened the usefulness of projects such as this latest one (see *Nature* **497,** 287; 2013).

The key, as always, is consent. The information at stake here is not genomic clues to future health risks — already the subject of fierce debate — but sensitive data on past and current medical conditions. What the government leaflet fails to highlight is the real threat to privacy and the possible consequences. Worse, the public-relations exercise carried out by the government to stress the programme's benefits has if anything increased the backlash from privacy campaigners, who are now highlighting the risks and urging people not to participate. An unfortunate false choice has been established, between scientific progress on one side and protection of privacy on the other.

The government did not initially intend even to allow individuals to opt out of having their data centralized in this way, which would have flown in the face of the most basic principles of privacy and informed consent. The leaflet now states, "You have a choice", but the government seems to have made it as difficult as possible for people to exercise that right. They must explicitly contact their local doctor to opt out — a requirement that seems a sure way to make certain that most won't bother, and so will be opted in by default. UK medical charities, including the Wellcome Trust, have launched their own advertising campaign in support of care.data, which, although it validly highlights many of the research opportunities of such big data, also fails to mention sufficiently prominently that an opt-out option exists, and indeed seems intended to try to reduce the number of people who opt out.

Maximizing the number of people entering the programme is clearly a noble goal. But one cannot help but get the uncomfortable impression that, in their enthusiasm to amass these data, the authorities are using sleight of hand and paying lip service to the principles of informed consent. Inconvenient as it may be, and even if it has some negative effects on the utility of the database, the opt-out option to care.data should be prominently displayed, and facilitated.

The public-relations campaigns also break the first rule of risk communication, which is to state clearly any potential, even if remote, risks. They are far too reassuring, for example, that people's personal data are in safe hands and will be well protected from abuse. Under the programme, personally identifiable data will be stored securely by the Health and Social Care Information Centre in Leeds, which will review requests for them. Most will be made available only after being pseudonymized — a process by which data are stripped of information that would otherwise easily allow identification of the data's personal provenance.

"The removal of identifiers is an essential step in protecting data, but it is far from foolproof."

This removal of identifiers is an essential step in protecting data, but it is far from foolproof, and a determined effort can often re-identify pseudonymized data. Furthermore, the Health and Social Care Information Centre's store of personally identifiable data is not immune from hacking or other intrusions. No doubt much thought has gone into protecting the data to high standards, but overly reassuring the population that its personal data are safe is an invitation to public disillusionment in the system down the road.

The potential gains for health authorities and researchers from patient-level data are immense. But both should be insisting on the spirit of informed consent — clearer, more-upfront information and greater visibility given to people's right to opt out would be a good start. ■

Cool heads needed

As cold weather rages, it is easy to forget the difference between weather and climate.

The United Kingdom had its stormiest December for 50 years last month, with harsh winds, downpours and flooding that extended into the new year. Across the Atlantic, North Americans found themselves hunkering indoors — or stuck in airports — as a mass of Arctic air delivered snow and record low temperatures that brought services to a halt. As tends to happen whenever the weather goes haywire, many wondered whether larger forces were at work; British Prime Minister David Cameron said he suspected that global warming was partly to blame for his country's suffering. It is a natural assumption, and scientists are actively engaging with the issue. But, as always, a little caution is in order.

On some level, most people understand the difference between climate and weather. Climate is the context: the accumulation of temperatures and precipitation trends that vary depending on location and season. Weather is what we experience, and extremes are part of the package. This was the message delivered by the UK Met Office, which pointed out that stormy conditions are more likely during winter months. Despite such assessments, however, people continually confound weather and climate in the heat — or cold — of the moment. Confusion seems unavoidable.

In the United States, the cold snap extended as far south as Florida, forcing thousands of flight cancellations at the height of the holiday season. Climate sceptics celebrated, apparently unable or unwilling to accept that even a warming planet experiences cold temperatures. A small cohort of scientists countered with arguments that global warming might in fact be contributing to the string of abnormally cold US winters in recent years. The argument is that rapid Arctic warming and melting sea ice are destabilizing the fast-flowing air current known as the polar jet stream, leading it to the kind of drunken meandering that can push Arctic air across North America — and deliver powerful storms to the United Kingdom. If that is true, Cameron may well have been right.

Evidence for the claim that global warming could be disrupting the jet stream is disputed. Similar weather events have happened in the past, and at least one review of the record suggests that nothing is amiss — at least nothing that scientists can pin down as obviously outside the normal year-to-year seasonal variations. This does not mean that climate change has no role, of course. It just means that we do not yet know. In the words of one climate modeller, until the models and the observations align, we ought to reserve judgement. As far as the public is concerned, there is little to do but dress appropriately, keep an open mind and let the science play out.

The same dynamic has been playing out in recent years with regard to the average global temperature, which has plateaued since 1998. At first blush, the global-warming 'hiatus' runs counter to the warming projected by climate models. Here again, climate sceptics have pounced, and some climate scientists have rightly begun to explore both the climate system and their models to sort out the apparent discrepancy. As reported on page 276, researchers are homing in on a potential explanation that ties the periodic warming and cooling of the eastern equatorial Pacific Ocean to global temperature trends. In particular, the cool phase of the Pacific Decadal Oscillation — which took hold in 1998, coinciding precisely with the hiatus — seems to drive heat into the ocean, effectively cooling the atmosphere.

"There are many ways to estimate the climate's likely response to greenhouse gases, and the evidence cuts both ways."

Plenty of questions remain. According to this theory, temperatures will rise anew when the eastern Pacific flips into its warm phase in the coming years. But how much warming should we expect when that happens? Exactly how sensitive is Earth's climate system to increasing atmospheric levels of greenhouse gases? Some have argued, in part on the basis of current temperature trends, that climate models tend to overestimate warming, which would be good news indeed if true. But there are many ways to estimate the climate's likely response to greenhouse gases, and the evidence cuts both ways.

Ultimately, the hiatus has provided an opportunity to better understand both the climate system and climate models. One lesson is that the climate, like day-to-day weather, has its ups and downs. Another is that the average global temperature — although a useful indicator — is not the only measure of how the climate changes. Scientists are still trying to work out what all of this means for the future, but if the past is any indication, we may have to live with a fair degree of uncertainty. From a policy perspective, little has changed. The range of potential impacts projected by climate models warrants much more aggressive action than has been initiated so far. ■

V is for vortex

An endangered species helps scientists to learn why migrating birds fly in a familiar formation.

The northern bald ibis (*Geronticus eremita*) was once such a widespread sight in the skies of north Africa that the bird was immortalized as an ancient Egyptian hieroglyph. The picture symbol denoted the word *akh*, which means 'to be resplendent, to shine'. Ibis populations are less resplendent today, with just a few hundred of the wild birds remaining, mainly in Morocco. They can still shine, however; a study of 14 northern bald ibises reported this week on page 399 offers the first experimental evidence that helps to resolve one of the great questions of the natural world: why do migrating birds often fly in an elegant V formation?

The obvious answer is that it saves energy. Just as the mass ranks of a peloton in a cycle race make life easier for riders, and as tight formations can save aircraft fuel, the signature shape of a flock of ibises or geese is assumed to make flight less of a flap — at least for the bulk of the birds that follow the leader. (That is another, less obvious, theory for the V shape: that the bird at the front is the best navigator.)

Some of the most influential research studies do little more than test whether the obvious answer to a question is the correct one. When it comes to bird flight, the validity of the obvious answer has, until now, been concealed by an obvious problem. Namely, that the equipment for monitoring the flight of wild birds tends to disappear over the horizon along with the bird to which it is attached. (Sensors that are able to relay the data tend to be too heavy for birds to carry.)

This is where the endangered plight of the northern bald ibis offered an opportunity to science. Several captive-breeding programmes exist, and a big part of preparing the birds for release is to teach them their traditional migration routes. Hand-reared ibises are trained to follow conservation experts who are inside a microlight aircraft. So, crucially, when these birds set off to fly in formation, they come back.

Steven Portugal, a researcher at the Royal Veterinary College in Hatfield, UK, used the training flights of ibises raised at a zoo in Vienna to test the benefits of formation flying. His team fitted the birds with lightweight data loggers that could measure both their body position and flapping movements.

The juvenile birds took a while to get into shape; a V formation is harder to achieve and maintain than it looks, and it looks pretty difficult. (RAF pilots told to fly in a tight V shape during the Second World War spent more time watching the position of the plane in front than scanning for enemy fighters.) Still, the 14 ibises did manage it for long enough for the scientists to accurately record both the distance between each bird and the timing of the creatures' wing flaps.

The results: when in formation, each bird was able to synchronize the flapping of its wings so that it could exploit the updraught created by the swirling vortex of air from the flapping wingtip of the bird in front. When the flock got it right, each following bird delayed its wingbeat by just enough to spread a wave of synchrony through each arm of the V. When they got it wrong and a following bird drifted directly behind the bird in front, the follower registered the problem and adjusted the timing of its flaps so that it did not become tangled in the powerful downdraught of the same vortex. For more, see the associated News & Views article on page 295. Or look up at the sky, and delight in the rare beauty of an obvious answer. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

Physiological data must remain confidential

Electronic devices that track our emotions, heart rate or brain waves should be regulated to protect individual privacy, says **Stephen Fairclough**.

How many people have beneath their television a shiny Xbox One? This new console is significant, and not just for its games and graphics. It is fitted with a camera that can monitor the heart rate of people sitting in the same room. The sensor is primarily designed for exercise games, allowing players to monitor heart changes during physical activity, but, in principle, the same type of system could monitor and pass on details of physiological responses to TV advertisements, horror movies or even party political broadcasts.

The Xbox One is the first consumer electronic hardware to permanently integrate a technology called physiological computing. Those of us who work in this field aim to transform the way that people use, control and interact with electronic devices in everyday life. But like all technology it has a darker side, and chief among the concerns is the potential infringement of privacy. The mass appeal of the Xbox One suggests the probable broad reach of such tools, and now is a good time to consider both the benefits and risks of the technology — particularly in the wake of the first international conference devoted to this topic, which took place last week in Lisbon (www.phycs.org).

Most people do not consider the human body to be a transmitter, but our nervous systems constantly generate data — from the first pulse of a fetal heartbeat to a dying breath. Physiological computing converts these data to control inputs for a computing system, using the signals as a proxy for the conventional keyboard and mouse. Brain–computer interfaces, for example, can already move a cursor on a screen in response to electrical fluctuations in the brain.

The same technology can monitor spontaneous activity from the brain and body to infer a computer user's emotional and cognitive state. For instance, moods such as anger or frustration can be detected from specific changes in cardiovascular activity and breathing patterns. And increased concentration on difficult mental tasks produces characteristic changes in brain activation that can be picked up in an electroencephalogram (EEG).

Scientists want to use these physiological changes to create technology that can respond to the circumstances and adjust conditions to improve the quality of the human–computer interaction. A desktop computer that can recognize frustration from cardiac data could be programmed to offer help or even to play soothing music; sensors in a phone could spot stress during a fraught journey in heavy traffic or poor weather and automatically divert all calls to voicemail. This scenario, in which software adapts in a proactive and implicit way to dynamic signals from the user, represents a radical departure from how we currently interact with computers.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/zgwoiz

INFORMATION GATHERED ON A PERSON'S PHYSIOLOGY SHOULD BE CONSIDERED TO BE OWNED BY THAT PERSON.

A good example is digital health, in which wireless devices and sensors can record physiological activity to offer a wealth of quantitative information about lifestyle and well-being. These data can reveal the impact of changes in exercise or diet on physiological markers such as cardiovascular activity. A colleague who wore a chest band non-stop for a year to monitor his heart rate learned, for example, how workload affected his sleeping patterns. This type of ambulatory measurement — and the cumulative gathering of information — delivers big data at the level of the individual.

Until now, the main barrier to developing this technology has been the scarcity of sensors that are both inconspicuous and capable of delivering high-quality data. But the field of wearable sensors is evolving at an extraordinary pace. The traditional image of the laboratory participant festooned in wires is being replaced by one in which discreet, ambulatory sensors stream data to mobile devices. The ubiquity of cameras on smartphones means that, with the right app, heart rate can be detected from the finger or even remotely from the face. As sensors improve, so too will their public acceptance. Their spread, in turn, will boost the quality of the data they can generate and the number of uses to which they can be applied. For example, continuous monitoring of EEGs using ambulatory equipment can reveal patterns of brain activity characteristic of epilepsy — useful information not only for individuals but also for the health-insurance industry. And these advances prompt questions such as: who owns the data? Who should be allowed to gather and store this information?

As a researcher, I would never monitor the physiology of a person in the lab or field without consent. But privacy concerns are real, and I think that most people would be more comfortable with this type of technology if protection or regulation were in place sooner rather than later. Advances in genomics and gene sequencing have raised legitimate concerns about the ability of third parties to covertly obtain and screen someone's DNA — taken perhaps from a used coffee cup to test for paternity when it is disputed, for instance. (UK law demands consent for such tests.) Similarly, the field of physiological computing needs to decide on rules and guidelines for researchers and others.

We are at the start of this debate, but there is one key point that should underpin all future discussions. Information gathered on a person's physiology should be considered to be owned by that person. The default position must be that these data should be confidential in the same way as medical records, for that is what they are. ■

Stephen Fairclough *is professor of psychophysiology at Liverpool John Moores University, UK, and blogs at physiologicalcomputing.net. e-mail: s.fairclough@ljmu.ac.uk*

RESEARCH HIGHLIGHTS

Selections from the scientific literature

MATERIALS

Warm carbon coat reduces friction

A coating material made of carbon reduces friction not just by providing a slippery surface, but also by keeping the points of contact warm.

Marcus Björling of Luleå University of Technology in Sweden and his team coated steel balls with 'diamond-like carbon' — a material in which the carbon atoms have a bonding pattern similar to that of diamond. They rolled the balls against a metal disk with an oil lubricant in between, and showed that the carbon coating acts as an insulator, lowering the viscosity of the lubricant and thus reducing the friction between the ball and the disk.

The findings could encourage the development of lubricant coatings made from insulating materials.

Tribol. Lett. **http://doi.org/qtm (2014)**

RAVELL CALL/DESERET NEWS/AP

GEOLOGY

Landslide triggered earthquakes

Large earthquakes often cause landslides, but in an unusual reversal, a massive landslide in a US copper mine in April 2013 might have resulted in a series of small earthquakes.

Kristine Pankow and her colleagues at the University of Utah in Salt Lake City describe two sequential rock avalanches at the Bingham Canyon mine (pictured) near Salt Lake City. Together, these events comprise probably the largest non-volcanic landslide to have occurred in North America in modern times. The proximity of the landslide to a seismic monitoring network produced extensive data, which showed that the avalanches had an estimated magnitude of about 2.5. In the days following the landslide, the sensors detected 16 more seismic events.

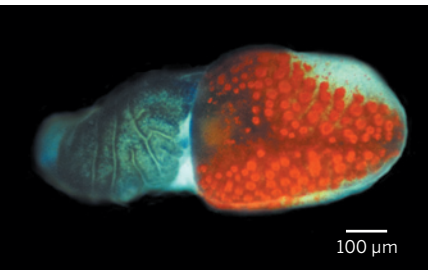
GSA Today 24, **4–9 (2014)**

ZOOLOGY

Starfish eyes see the light

Starfish can use eyes at the tips of their arms to find their way home.

Most sea-star species have compound eyes on their arms, but there has been no direct evidence that the eyes enable vision. In field experiments, Anders Garm at the University of Copenhagen and Dan-Eric Nilsson at Lund University in Sweden displaced blue starfish (*Linckia laevigata*) from the coral reefs they inhabit off the coast of Japan. The starfish were able to navigate their way back to the reef from distances of up to two metres, an ability that was lost when the researchers surgically removed the animals' eyes.



DAN-ERIC NILSSON

The distribution and shape of the eyes (**pictured**) and the arrangement of their light-sensing cells suggest that the starfish can recognize the reef from only relatively short distances. This might help the creatures to stay close to home, the researchers say.

Proc. R. Soc. B 281, **20133011 (2014)**

MOLECULAR BIOLOGY

RNA retrieved from intact tissue

A technique can snatch RNA out of a single cell in live, intact tissue, revealing the genes being expressed by that cell.

Other methods of single-cell RNA retrieval tend to disrupt the surrounding tissue before the RNA is captured. James Eberwine of the University of Pennsylvania in Philadelphia and his colleagues created a molecule called a TIVA tag that penetrates cells.

When hit with a beam of light, the tag binds to RNA that is being made from its DNA template. By adding the tag to human and mouse brain tissue and shining a laser on individual cells, the researchers could activate the tag, retrieve the RNA and sequence it.

The approach could reveal how a cell's natural micro-environment affects its activity.

Nature Meth. **http://dx.doi.org/10.1038/nmeth.2804 (2014)**

PALAEONTOLOGY

Trilobites ventured beyond the ocean

Contrary to their reputation as denizens of the open ocean, the extinct creatures known

as trilobites may have crawled up on to tidal flats.

Fossils found by Gabriela Mángano at Canada's University of Saskatchewan and her colleagues in rock deposits from ancient tidal flats reveal that trilobites took to the land during the Cambrian explosion some 540 million years ago, when the number of animal species increased drastically.

The team discovered the fossils along with fossilized tracks in rocks from the US Appalachian Mountains. The rocks also showed signs of cracks from periodic drying, hinting that they originated in a tidal flat zone.

The finding supports the idea that terrestrial creatures evolved from marine rather than freshwater ancestors. Intertidal zones could have provided food or safe havens for these animals, the researchers suggest.

Geology **http://doi.org/qnq (2013)**

METABOLISM

How exercise benefits the body

A small molecule produced by muscles in response to exercise boosts metabolism in other tissues.

Robert Gerszten of Massachusetts General Hospital in Boston and his colleagues discovered the molecule, BAIBA, when they forced muscle cells to express the metabolic regulator PGC-1α — levels of which increase with exercise. BAIBA levels increased in exercising mice. In animals treated with BAIBA, white fat tissue showed greater expression of genes linked to calorie burning, and the mice gained less weight and had better glucose metabolism than untreated mice.

The researchers also found an inverse association in humans between BAIBA levels and heart-disease risk factors — people with more BAIBA in their blood also had decreased cholesterol levels and less insulin resistance, for instance. BAIBA could be a target for drugs that treat diabetes and other metabolic disorders, the authors say.

Cell Metab. 19, **96–108 (2014)**

CLIMATE CHANGE

Past warmth drives glacial melting

The world's glaciers will probably continue to shrink over the next decades, irrespective of the magnitude of future warming.

Ben Marzeion at the University of Innsbruck in Austria and his colleagues ran a global glacier model using various twenty-first-century scenarios for greenhouse-gas concentration. They found only small changes in the loss of mass from glaciers under greatly different climate-change conditions.

Projected glacier melting this century is essentially a delayed response to climate changes in the twentieth century, the authors say. The thinning and retreat of glaciers from low-lying areas make them less sensitive to future warmer temperatures, they conclude.

Cryosphere 8, **59–71 (2014)**

MICROBIOLOGY

Marine bacteria shed tiny sacs

The most abundant photosynthetic bacterium in the oceans casts off many minute pieces of itself every day, amounting collectively to tonnes of material that potentially influences the global carbon cycle.

Many bacterial species release membrane-bound sacs called vesicles, which have not been well studied in natural ecosystems. Sallie Chisholm, Steven Biller and their colleagues at the Massachusetts Institute of Technology in Cambridge discovered vesicles in laboratory cultures of the microbe *Prochlorococcus*, and in samples from the Atlantic Ocean.

Analysis revealed that the laboratory vesicles contained proteins, DNA and RNA, and that each *Prochlorococcus* produced two to five vesicles per generation. The authors estimated that the *Prochlorococcus* sacs could be contributing 10^4 tonnes or more of fixed carbon to the ocean carbon cycle each day. Vesicles might serve to decoy attacking viruses away from the bacterium and aid in gene transfer.

Science 343, **183–186 (2014)**

OCEANOGRAPHY

Sea-level swings get more extreme

The seasonal rise and fall in sea level along the US Gulf coast has grown more pronounced since the 1990s compared with earlier decades, probably because of warmer summers and colder winters.

Thomas Wahl and his colleagues at the University of South Florida in St Petersburg compared sea-level measurements collected between 1900 and 2011 with atmospheric data for the Gulf of Mexico coastline. They found that typical differences in sea level between summer and winter have increased during the past two decades.

Higher summer sea levels could increase the chances of hurricane-related flooding, and even slight changes in both summer and winter sea levels may affect sensitive ecosystems, the authors say.

Geophys. Res. Lett. **http://doi.org/qtd (2014)**

COMMUNITY CHOICE

The most viewed papers in science

PHYSICS

Why penguins do the wave

HIGHLY READ
on iopscience.iop.org
9 Dec–8 Jan

Physicists have explained how waves of coordinated motion sweep through huddles of male emperor penguins (*Aptenodytes forsteri*; pictured) as they try to keep warm while incubating eggs in the Antarctic.

Daniel Zitterbart and Richard Gerum at the University of Erlangen-Nuremberg, Germany, and their colleagues analysed video recordings of penguin huddles and built a mathematical model to study the waves. The authors found that any penguin taking a step of two centimetres or more within a densely packed huddle can trigger ripples of disturbance as nearby penguins readjust to keep close (but not too close) to each other.

The movements were similar to those seen in traffic jams in which waves begin at the front of the queue and travel backwards. However, in penguin huddles, waves can move in multiple directions from any location.

New J. Phys. 15, **125022 (2013)**

FRANS LANTING/CORBIS

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Power-plant rules

The US Environmental Protection Agency (EPA) published a controversial rule on 8 January governing greenhouse-gas emissions from new power plants. Advanced natural-gas power plants are poised to meet the standards, but the rule would effectively require new coal-fired plants to capture and sequester about 40% of their emissions — a feat that many industry officials have criticized as technologically infeasible. The EPA will now accept public comments on the rule while working on a second standard governing existing power plants.

Polio progress

On 13 January, India marked its third consecutive year without a case of polio, clearing the way for the World Health Organization to certify the southeast Asia region as polio-free. The achievement is a major milestone for India, where high population density and poor sanitation had enabled the poliomyelitis virus to spread. Pakistan, Afghanistan and Nigeria remain the only countries never to interrupt transmission of polio, and the virus re-emerged last year in war-torn Syria and the Horn of Africa.

NASA/ESA/J. LOTZ, M. MOUNTAIN, A. KOEKEMOER & THE HFF TEAM (STSCI)

Super-distant galaxies glimpsed

Astronomers unveiled pictures of the deepest galaxy cluster ever imaged at the annual meeting of the American Astronomical Society in National Harbor, Maryland, which ended on 9 January. The images from NASA's Hubble Space Telescope are part of the Frontier Fields programme, which harnesses the phenomenon of 'gravitational lensing' (see *Nature* **497,** 554–556; 2013). The tremendous gravity of large foreground clusters — in this case, Abell 2744 (pictured) — distorts space, enhancing the visibility of more-distant galaxies. Abell 2744, which shows hundreds of galaxies as they looked 3.5 billion years ago, produced gravitational lensing that allowed scientists to see background galaxies from more than 12 billion years ago. Some of the objects captured are 10–20 times fainter than any galaxies previously observed.

BUSINESS

Novartis woes

The Japanese health ministry filed a criminal complaint on 9 January against Swiss pharmaceutical firm Novartis, claiming that the Basel-based company had exaggerated the benefits of Diovan (valsartan), used to lower blood pressure. Advertisements for the best-selling drug relied on studies showing that it also reduced the risk of stroke and heart attack. But by early 2013, some related papers had been retracted over flawed data and analysis (see go.nature.com/hdfzbi for more). Novartis's Japan unit has acknowledged the complaint on its website.

Cloud computing

IBM has invested US$1 billion in the IBM Watson Group, a business unit to commercialize artificial-intelligence applications that can be accessed by customers remotely. Located in New York City, the group is based around Watson, the computer that famously won against human competitors in a quiz show in 2011 (see go.nature.com/u783dz). Watson has since been farmed out to analyse large data sets, and is being tested at the Memorial Sloan-Kettering Cancer Center in New York City to help physicians decide how to diagnose and treat patients.

RESEARCH

Space station stays

As space-agency leaders from around the world gathered in Washington DC to discuss the future of space exploration, the US White House approved operations aboard the International Space Station until at least 2024, extending the previous 2020 end date. In a joint announcement on 8 January, Charles Bolden, the head of NASA, and President Barack Obama's chief science adviser John Holdren said that the decision will enable the continuation of short- and long-term research, including planned human missions to an asteroid and Mars.

EVENTS

Welcome break-up

Two ships that had been stranded in thick ice off Antarctica reported finally heading for open waters on 8 January. The Russian vessel *Akademik Shokalskiy* got stuck near Commonwealth Bay on 24 December during

a research voyage, and the Chinese icebreaker *Xue Long* became trapped during a rescue attempt (see *Nature* **505,** 133; 2014). Shifting weather conditions allowed the ships to break free. See pages 270 and 291 for more.

Industrial blast

An explosion at a chemical plant in Japan on 9 January killed 5 and injured 12. The blast at Mitsubishi Materials in Yokkaichi, about 300 kilometres west of Tokyo, occurred while workers cleaned a tank used to cool gas during the manufacture of silicon. The company is still investigating the cause of the explosion, but some reports suggest that residual chlorine or hydrogen in the tank may have reacted with air.

Animal activism

Anti-terrorism police in Italy are investigating the targeting of four scientists at the University of Milan who use animals in their research. In the early hours of 7 January, activists posted flyers and graffiti around the neighbourhoods where the scientists live, giving their names, photographs, addresses and telephone numbers. The flyers described the researchers as torturers and murderers, and exhorted readers to harass the scientists by phone.

PEOPLE

DANIEL LYNCH/REX FEATURES

Engineering leader

The Royal Academy of Engineering in London announced the nomination of Ann Dowling (**pictured**) as its first female president on 9 January. Dowling, who heads the department of engineering at the University of Cambridge, UK, became the department's first female professor in 1993. Her research focuses on achieving efficient, low-emission combustion and developing low-noise vehicles; she has received royal honours for her services to mechanical engineering and science. The academy is scheduled to elect Dowling formally in September.

AIDS chief

On 9 January, US President Barack Obama nominated physician Deborah Birx to coordinate the country's global AIDS efforts and administer the US President's Emergency Plan for AIDS Relief (PEPFAR). Birx currently heads the AIDS programme at the US Centers for Disease Control and Prevention in Atlanta, Georgia. The PEPFAR programme, which receives about US$6 billion per year to distribute antiretroviral drugs and medical care in countries affected by AIDS (see *Nature* **457,** 254–256; 2009), received a five-year reauthorization in December 2013.

USGS chief

US President Barack Obama announced on 9 January that he had nominated Suzette Kimball to lead the US Geological Survey. Kimball, a former deputy-director for the agency, has served as acting director since February 2013, when previous head Marcia McNutt resigned (see go.nature.com/eyn4uu). Kimball has also worked as an assistant professor of environmental sciences at the University of Virginia in Charlottesville, and as a regional chief scientist for the US National Park Service.

Smithsonian head

Plant pathologist Eva Pell announced on 8 January that she will step down as undersecretary for science at the Smithsonian Institution in Washington DC, less than four months after Wayne Clough announced plans to retire as the institution's leader (see *Nature* **501,** 467; 2013). Since 2010, Pell has headed the Smithsonian's science museums, its nine research centres and the National Zoo. Previously, she was senior vice-president for research and graduate-school dean at Pennsylvania State University in University Park.

COMING UP

20 JANUARY
The European Space Agency's Rosetta spacecraft comes out of hibernation in preparation for reaching its destination later this year: the comet 67P/Churyumov–Gerasimenko. The mission will be the first to land a probe on a comet's surface. See page 269 for more.
go.nature.com/ivxmmo

22 JANUARY
The European Union is set to unveil a package of long-term climate and energy goals and proposals.

FUNDING

Pesticide risks

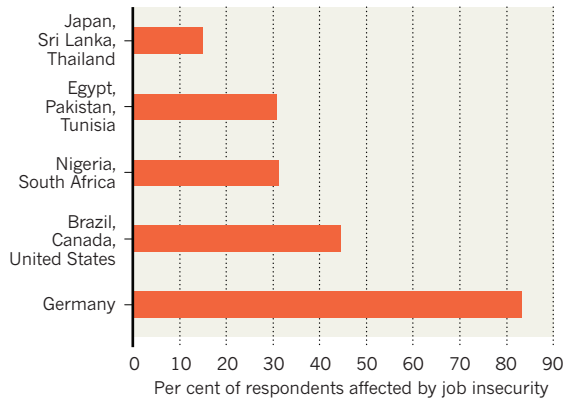
On 8 January, the US Environmental Protection Agency announced the award of nearly US$500,000 in grants for research to reduce the risks of pesticides, especially to bees. Scientists at Pennsylvania State University in University Park received funding to study alternatives to treating seeds with neonicotinoids, a class of pesticide linked to declines in bee populations (see *Nature* **496,** 408; 2013). At Louisiana State University in Baton Rouge, researchers will assess the long-term risks to bees from chemicals used in large-scale mosquito-abatement programmes.

TREND WATCH

Career barriers for young scientists vary widely across different regions of the globe, according to a survey of researchers in 12 countries that will be published this month by the Global Young Academy in Berlin. A large fraction of respondents in Germany reported problems with job insecurity, compared with other regions (see chart). More researchers in Egypt, Pakistan and Tunisia cited political instability as a career barrier than their peers around the world.

SHAKY FUTURES

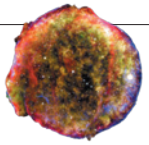
Job insecurity causes problems for early-career scientists around the world, especially in Germany, according to a recent survey.



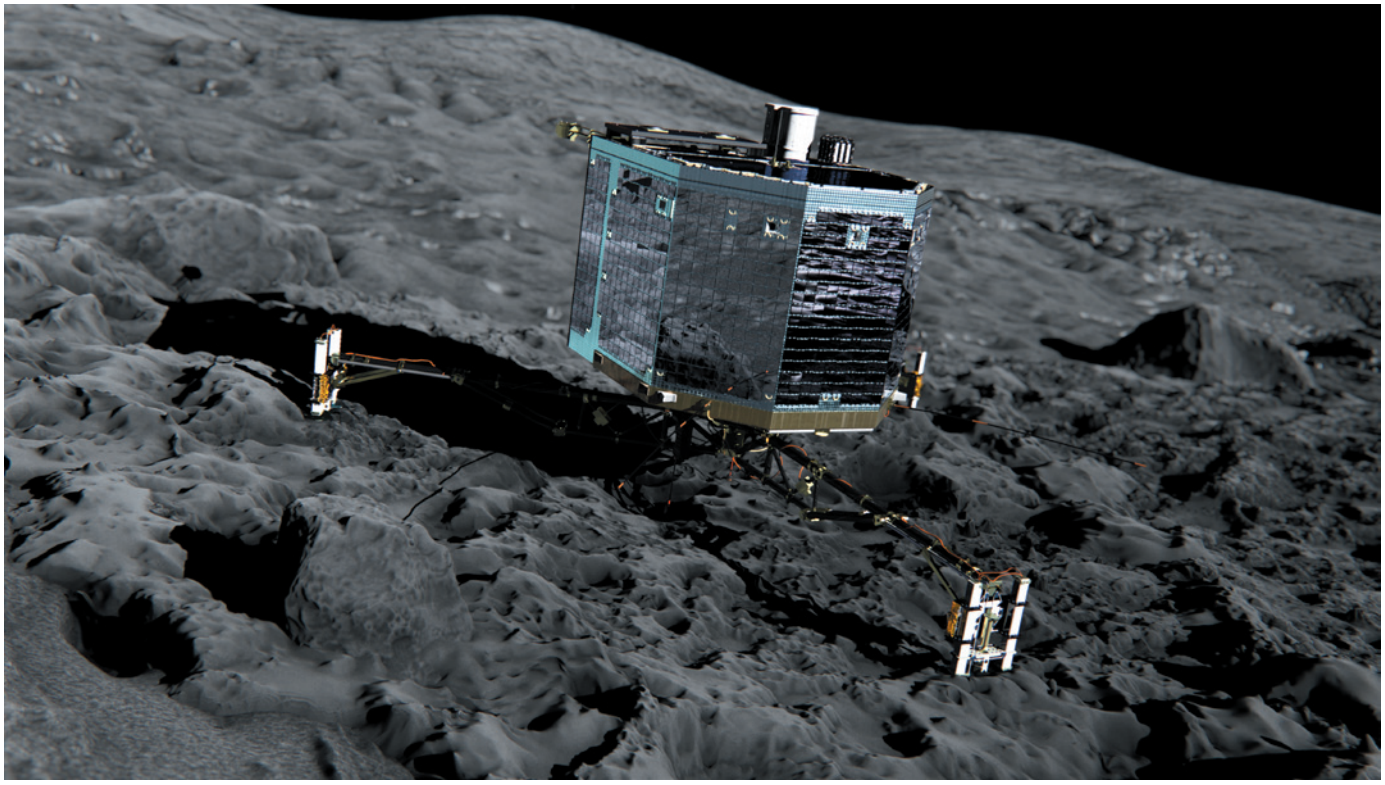
SOURCE: GLOBAL YOUNG ACADEMY

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS



ATG MEDIALAB/ESA



In November 2014, the Philae lander will be the first to try to land on the surface of a comet, 67P/Churyumov–Gerasimenko, as shown in this artist's impression.

SPACE

Comet craft ready to wake

Stakes high as European Space Agency waits for Rosetta orbiter to come out of hibernation.

BY ELIZABETH GIBNEY

Space scientists are used to moments of high tension. They often have just one chance to get things right, and experiments can hinge on the success of equipment that may be millions of kilometres away. So there will be considerable anxiety on 20 January at the European Space Agency (ESA) when the comet-hunting spacecraft Rosetta is due to stir after almost three years of hibernation.

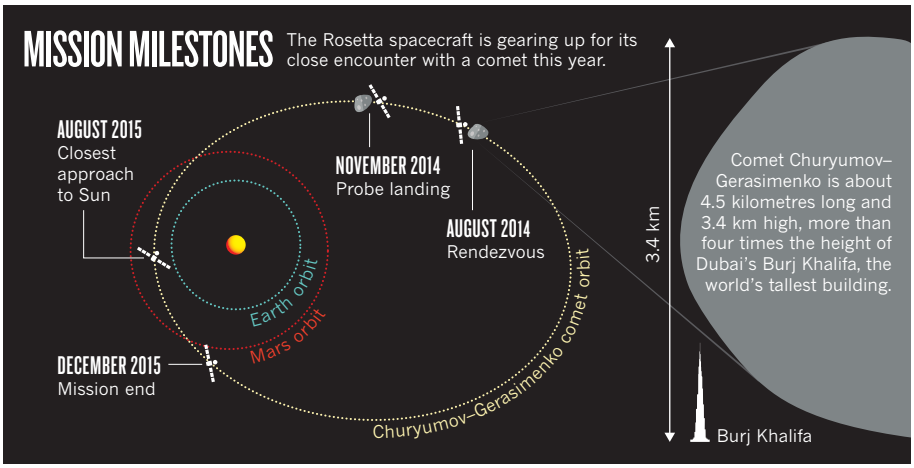
With Rosetta now some 800 million kilometres from Earth, and rapidly approaching its target — comet 67P/Churyumov–Gerasimenko — the first sign that things are going to plan on Rosetta will be the activation of a pre-set alarm. This will trigger a series of automated events that should see the craft's components warmed up, its spin corrected with thrusters and an antenna pointed at Earth to begin communications. There will be an anxious wait.

"I can tell you, you sweat like hell," says Claudio Sollazzo, operations manager of ESA's 2005 Huygens mission to Saturn's moon Titan, who endured about two weeks of radio silence after the probe was released from the Cassini orbiter. "With every second of delay, immediately you say, 'OK — something has gone wrong.' You can know you've done everything right, but it's always human nature to believe something bad has happened."

Once Rosetta is awake, the stakes will get higher. Later this year, mission controllers will attempt to observe the comet up close and land the craft's probe, Philae — the first time a comet landing has been attempted. Both objectives aim to deliver vital data about the formation of the Solar System and life on our planet.

The €1-billion (US$1.4-billion) mission has been hurtling through the Solar System since 2004. Its 7-billion-kilometre journey has seen it fly past asteroids, and slingshot around Earth and Mars before it was put into hibernation to save energy in June 2011.

Rosetta will arrive at its target in August, when it will map the comet's surface, gravity, shape and rotation to find a suitable site to dispatch the metre-cubed Philae lander (see ▶



SOURCE: ESA

'Mission milestones'). "We really know very little about this object," says Andrea Accomazzo, Rosetta's spacecraft-operations manager.

ESA plans to release Philae in November, before the comet — already spewing gas and dust — gets too close to the Sun, when it will start to produce more debris. Rosetta, which will ordinarily orbit at up to 100 kilometres from the comet, will descend to about 4 kilometres away for the landing, and will drop its instrument-laden probe unpowered into a zone about one kilometre squared. With just one shot at the landing, which will be automated because of a 30-minute delay in communications with Earth, the operation will be risky. "Even though we will have chosen the best possible site, we'll still need some luck to land on a safe part," says Mark McCaughrean, senior scientific adviser at the ESA directorate of science and robotic exploration. Rosetta will circle the comet at the equivalent of walking pace, albeit as both hurtle along at 20 kilometres per second.

Once Philae has secured itself on the surface using a harpoon, its batteries will last a few days; after this it must survive on solar power, which will be difficult once its panels are covered in dust. "We really don't know how long it will last," says McCaughrean. Although the emphasis will be on the lander's performance, just getting the orbiting spacecraft and its complex instruments into position around the comet will be a big success, he says. "If we put the lander on the surface, that's the cherry on the cake."

Comets are primitive objects. Their gas, dust and organic molecules have changed little since their creation along with the Solar System 4.6 billion years ago. Scientists think that they hold strong clues about its origins. Comets are also thought to have delivered a large fraction of Earth's water, and possibly amino acids, the building blocks of life.

Assuming all goes to plan, cameras and sensors on the orbiter will examine the comet in detail over the next year, while spectrometers analyse the chemistry of its dust. The lander will look at the comet's surface composition and structure. Both will assess the ratio of normal to heavy water — formed from the hydrogen isotope deuterium — to see if comet ice matches Earth's water signature, and their instruments will hunt for the complex organic molecules that are needed to assemble primitive life.

Uwe Meierhenrich, an analytical chemist at the University of Nice Sophia Antipolis in France, will be anxiously waiting for the results. He is co-investigator of Philae's COSAC (Cometary Sampling and Composition) experiment, which will analyse materials from around 20 centimetres beneath the comet's surface. These might include organic materials that do not vaporize and never form part of the gas tail that astronomers can study from Earth.

COSAC will also measure the chirality (or 'handedness') of any detected amino acids, something that is impossible through remote observation and has never been done before on comet samples, says Meierhenrich. On Earth, amino acids in proteins are left-handed, so finding a predominance of left-handed molecules on the comet would add weight to theories that such a cosmic traveller seeded life on Earth by providing essential basic ingredients. ■

ANTARCTIC EXPLORATION

Researchers question rescued polar expedition

Australian Antarctic Division says it did not approve research strategy of stricken mission.

BY ALEXANDRA WITZE

The Australian Antarctic Division (AAD) has added its voice to the growing criticism of a stricken private polar expedition by challenging claims that it approved the research element of the trip.

On page 291 of this issue, Nick Gales, chief scientist of the AAD, which is based in Kingston, Tasmania, responds to an earlier *Nature* column by expedition head Chris Turney of the University of New South Wales (see *Nature* **505,** 133; 2014). Turney's Australasian Antarctic Expedition aimed to retrace the steps of explorer Douglas Mawson, who led an outing a century ago. But members of Turney's expedition had to be rescued from their ship, the *MV Akademik Shokalskiy*, after it became trapped in ice at Christmas, adding fuel to a debate about the merits of such privately funded trips.

Gales challenges Turney's implicit suggestion that the AAD had approved the expedition's science plan. The AAD did not formally review the research strategy, Gales notes, but had issued the permits required for Turney's group to visit the region in which it got stuck. "It's an important distinction for us," says Gales. He adds that the expedition's rescue has delayed several projects in Australia's national Antarctic programme that have been many years in the planning.

Other polar scientists have criticized the nine-point science plan laid out on the expedition's website. The plan "could be interpreted as delivering outcomes that I believe are not possible from this single voyage", says Richard Coleman, deputy director of the Institute for Marine and Antarctic Studies at the University of Tasmania in Hobart, Australia. For instance, the plan's first bullet point says that the expedition aims to "gain new insights into the circulation of the Southern Ocean and its impact on the global carbon cycle". A single trip could provide only

Members of the Australasian Antarctic Expedition are rescued by a Chinese helicopter after their ship, the *MV Akademik Shokalskiy*, became trapped in ice.

ANDREW PEACOCK/FOOTLOOSEPHOTOGRAPHY.COM/EPA/CORBIS

limited glimpses into this massive question, says Coleman, who sits on the committee that evaluates research proposals for the AAD.

Turney, Gales and others all agree that the scientific value of the expedition will be measured in the quality of peer-reviewed science it produces. Yet the expedition's fate has sharpened the long-standing tension between government polar-research programmes, which typically follow strategic plans devised through many rounds of peer review, and private voyages, which occasionally have science objectives along with other goals such as tourism or raising environmental awareness.

Turney says that he regrets any confusion over the AAD's involvement in the expedition. "At no stage did I intend to convey the impression that the [expedition] projects had been subject to the competitive peer-reviewed process required for participants in the formal Australian Antarctic programme," he says.

The Russian-registered *Shokalskiy* became trapped on 24 December in thick ice in the Commonwealth Bay region. The vessel's captain put out a distress call, and Australian, Chinese, French and US ships interrupted their schedules and came to help.

On 2 January, the 52 scientists, students, educators and journalists aboard the *Shokalskiy* evacuated to the Australian vessel *Aurora Australis* aboard a helicopter provided by the Chinese icebreaker *Xue Long*, which had also become stuck in ice while attempting to reach the stricken *Shokalskiy*. The complex maritime emergency ended a few days later when the *Xue Long* and the *Shokalskiy* both managed to extricate themselves. The *Xue Long* continued on its way to scout the Ross Sea region for the site of a future Antarctic base, China's fifth.

The *Australis* travelled to Australia's Casey Antarctic base, where it had been working before the rescue, with its new passengers. The diversion has put it behind schedule, delaying the resupply of several of Australia's polar-research stations. A Casey-based project to study ocean acidification did not receive the diving equipment it needs to scout sites for a carbon dioxide enrichment experiment to be carried out this year, among other delays.

"We are now faced with the difficult situation of how to salvage our own project."

"The reduction in available field time for this season has set us back significantly and we will not achieve all of our goals for this summer," says Donna Roberts, the acidification project's chief investigator and a senior research fellow at the Antarctic Climate and Ecosystems Cooperative Research Centre in Hobart. "We are now faced with the difficult situation of how to salvage our own project."

Like other governments involved in Antarctic research, Australia has a long-term strategic science plan built around testable hypotheses to answer key questions, says Mahlon Kennicutt, a Texas-based oceanographer and former president of the Scientific Committee for Antarctic Research. "Those that have invested great time and energy to pass the high bar of national funding see their programmes being jeopardized by those that might be perceived to have circumvented the system," he says.

But expedition participant Janet Wilmshurst, a palaeoecologist at Landcare Research in Lincoln, New Zealand, argues that the expedition has brought back important science. Before reaching Antarctica, it explored the subantarctic islands of New Zealand. Wilmshurst led a team that gathered peat cores, leaf and soil litter and other samples to study environmental change.

"It was incredibly productive for us," she says. For instance, three new peat cores from the Snares Islands will be the first from the island group to be analysed using modern techniques, providing a glimpse into climate history at a key location where the northern Southern Ocean meets a subtropical front.

Sorting out the long-term research effects of the *Shokalskiy* rescue may take some time. "My biggest concern is that of the reputation for Antarctic science," says Patrick Quilty, a geologist at the University of Tasmania who has worked in Antarctica for nearly five decades. "There are ramifications." ■

LES GIBBON/ALAMY

MORE ONLINE

VIDEO OF THE WEEK

Freshwater fish jumps to catch bird out of the air go.nature.com/vc9kww

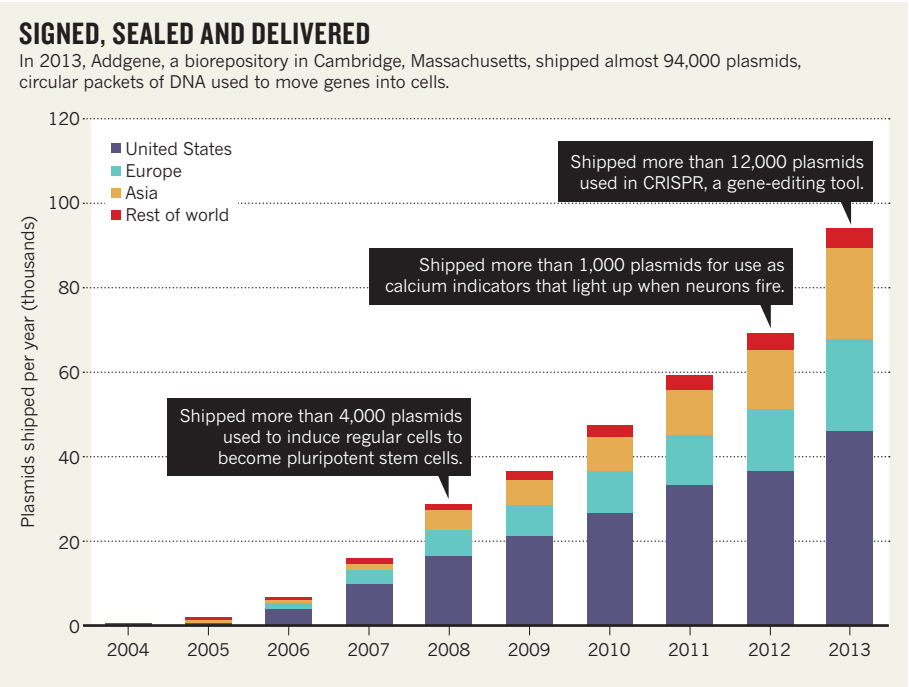
MORE NEWS

- Ground shifted 50 centimetres before Icelandic volcano erupted go.nature.com/mavw3u
- Nineteenth-century strain of cholera is rarely seen today go.nature.com/yaiol5
- More than 180 fish species are biofluorescent go.nature.com/iswnnj

NATURE PODCAST

Why birds fly in V formation, old trees sequester more carbon, and the FDA's decision on 23andMe nature.com/nature/podcast

ADDGENE



RESEARCH COMMUNITY

Repositories share key research tools

But some biological resource centres face funding issues.

BY MONYA BAKER

It was born out of the frustration of a graduate student trying to finish her thesis. In 2002, Melina Fan, then at Harvard Medical School in Boston, Massachusetts, found that the genes she needed for her experiments on metabolism were in other scientists' freezers. She wrote to more than a dozen labs to request certain plasmids — circular packets of DNA that scientists use to shuttle genes into cells. Only about half responded, and some plasmids took months to arrive. When they did, some contained the wrong genes.

"It wasn't that the scientists didn't want to share," says Fan. "It was just that the logistics were difficult."

So in January 2004, along with her husband and her brother, Fan founded Addgene, a repository in Cambridge, Massachusetts, in which scientists can deposit plasmids for free and order them for US$65 apiece. In its first year, Addgene shipped only one plasmid, containing a gene that Fan had worked on as a graduate student. Last year, the non-profit organization sent out more than 90,000 plasmids, ranging from specialized protein-tagging tools to empty DNA rings that can be filled with select genes.

Addgene's growth over its ten years demonstrates the extent to which biological resource centres (BRCs) have become crucial middlemen, enabling the sharing of tools that scientists say they are willing to distribute but often fail to. Where cell lines or genetic strains of model animals were once shared through unreliable, ad hoc networks, BRCs have allowed biological tools to be procured easily online.

"We are preserving the samples so they will be available to scientists in the future."

"BRCs are suppliers of public goods that are essential to supporting the rate of scientific progress," says Jeffrey Furman, an economist at Boston University in Massachusetts, who reported in 2011 that depositing cell lines, microbe strains and other items in BRCs could boost the citation rates of articles associated with the materials by 57–135%. And yet some are struggling to obtain the funding they need to survive.

The earliest BRCs have long, venerable histories. The American Type Culture Collection (ATCC) in Manassas, Virginia, established in 1925, holds more than 1 million genetic constructs, microbes and cell lines. And the Jackson Laboratory in Bar Harbor, Maine, established in 1929, maintains a collection of 7,000 strains of inbred and genetically modified mice.

Despite Addgene's relative youth, molecular biologists have flocked to its services. Addgene shipments reveal hot scientific trends such as the use of genetic-editing tool CRISPR. Last year, Addgene shipped more than 12,000 plasmids containing CRISPR tools (see 'Signed, sealed and delivered').

But although the repositories are valued, maintaining them can be challenging, especially if they hold expensive materials that are studied by few researchers, says Kevin McCluskey, curator of the Fungal Genetics Stock Center at the University of Missouri–Kansas City. The collection of more than 25,000 fungal strains is supported by a US National Science Foundation grant that has not been renewed.

"They want us to become self-sustainable, but then you have to raise your fees so much that you impact the ability of people to use your resources," McCluskey says. He notes that his collection charges a mere $30 per strain to academic labs, whereas those distributed by the ATCC can cost ten times as much. (A spokesperson for the ATCC says that the cost difference is justified by the breadth and quality of its collection.)

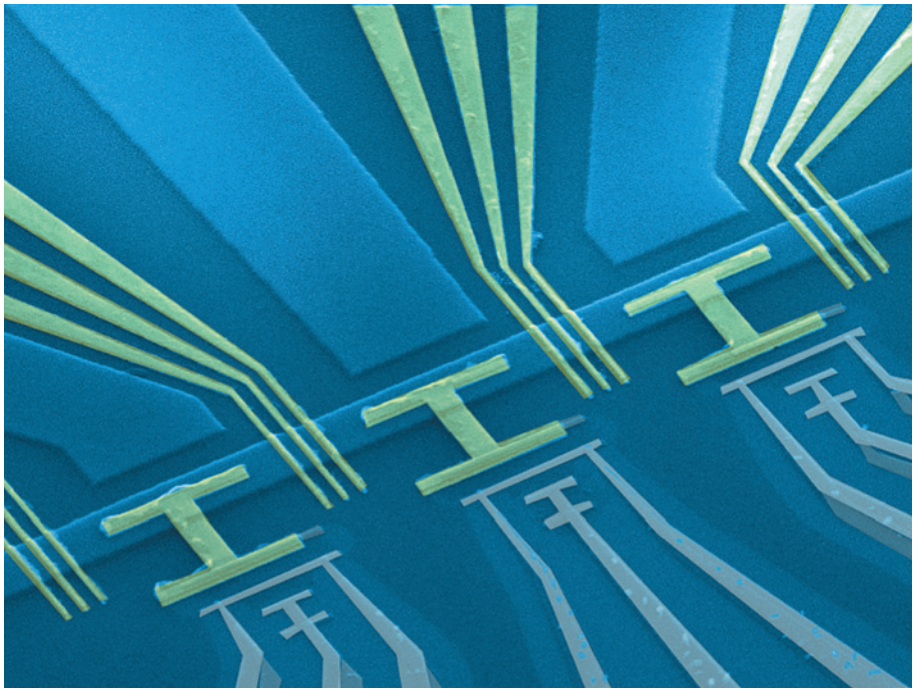
Changes in technology can upset the apple cart, too. Pieter de Jong, director of the BACPAC Resources Center at the Children's Hospital Oakland Research Institute in California, built up a diverse collection of artificial chromosomes to store and copy large chunks of high-quality DNA for sequencing projects elsewhere. At one point, the collection was used in the Human Genome Project.

But the rise of faster, cheaper sequencing has decreased demand for artificial chromosomes, and user fees no longer cover maintenance, says de Jong. Last year, he discarded backup materials and halved the number of storage freezers he keeps to cut costs. Next, he may need to prune unique material that is seldom requested, such as his platypus DNA collection.

But what is seldom requested today can be invaluable tomorrow. A heat-stable enzyme used in the 1980s to develop the polymerase chain reaction, an essential DNA-copying technique, came from bacteria deposited in the ATCC in the 1960s, with no consideration of its practical relevance.

At Addgene, about one-third of the plasmids are never requested, but the fees from the others cover the cost of maintaining them, says Fan. "We are preserving the samples so they will be available to scientists in the future." ■

PTB



A semiconductor device, seen in a scanning electron micrograph, can measure the flow of single electrons.

METROLOGY

Ampere to get rational redefinition

Single-electron flow measured in bid to overhaul SI base unit.

BY EUGENIE SAMUEL REICH

Physicists have tracked electrons crossing a semiconductor chip one at a time — an experiment that should at last enable a rational definition of the ampere, the unit of electrical current.

At present, an ampere is defined as the amount of charge flowing per second through two infinitely long wires one metre apart, such that the wires attract each other with a force of 2×10^{-7} newtons per metre of length. That definition, adopted in 1948 and based on a thought experiment that can at best be approximated in the laboratory, is clumsy — almost as much of an embarrassment as the definition of the kilogram, which relies on the fluctuating mass of a 125-year-old platinum-and-iridium cylinder stored at the International Bureau of Weights and Measures (BIPM) in Paris.

The new approach, described in a paper[1] posted onto the arXiv server on 19 December, would redefine the amp on the basis of *e*, a physical constant representing the charge of an electron. Metrologists have long sought such a rational definition. "It's an enormously challenging thing to try and do and it's quite an important paper," says Stephen Giblin, a physicist at the National Physical Laboratory in Teddington, UK.

The result will find favour at a meeting of the BIPM's General Conference on Weights and Measures in November. There, metrologists will discuss a proposal to redefine the ampere, the kilogram and two other standard (SI) units — the mole and the kelvin — in terms of the physical constants *e*, Planck's constant, Avogadro's constant and Boltzmann's constant.

Two other basic SI units, the metre and the second, have already been redefined in terms of two constants: the speed of light and the frequency at which electrons in caesium atoms transition between energy levels.

In the amp experiment[1], physicist Hans Schumacher of the Federal Institute of Physical and Technical Affairs (PTB) in Braunschweig, Germany, and his colleagues made use of a single-electron pump, a device in which voltage pulses prompt electrons to quantum-mechanically tunnel across barriers one at a time. The researchers tracked the paths of individual electrons by detecting changes in the electrical charge stored at points between ▶

the barriers. Primitive electron pumps have existed since 1990, but this is the first time that changes in charge have been detected for each hop of an electron.

The pump transferred just a few dozen electrons per second — slow enough to permit precision measurement and thus to provide proof of principle for redefining the amp. But this is only a first step: the set-up would not be practical for calibrating current-measuring ammeters, which need to run at higher currents. The ultimate goal is to create a '*mise en pratique*' — a standard-setting experiment that can be reproduced in any lab to calibrate measurements of current precisely — so the race is now on to combine Schumacher's validation method with a higher-current pump.

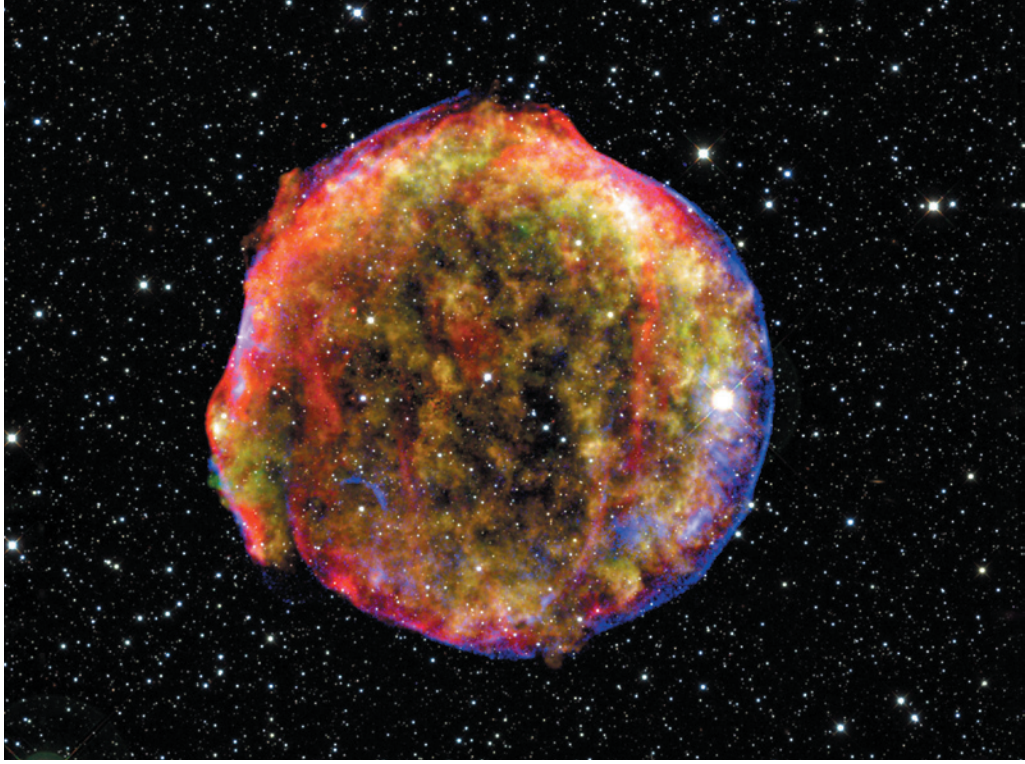
"It's impossible to say which is the winning concept."

In 2012, Giblin pioneered a semiconductor single-electron pump that transferred nearly one billion electrons per second (ref. 2), but he could not track them one by one. Other types of pump include turnstiles, in which electrons tunnel between superconducting wires, and tunnel junctions, in which electrons tunnel between aluminium islands separated by layers of insulating oxide. But these must also be operated at relatively low currents to track single electrons. "It's impossible to say which is the winning concept," says Jukka Pekola, a physicist at Aalto University in Espoo, Finland, who reviewed approaches to redefining the amp in 2013 (ref. 3).

Still, the amp is in good shape for the November meeting, as is the kelvin. The charge of the electron and Boltzmann's constant have both been measured precisely, so both units are ready to be redefined today, says François Piquemal, a physicist at the National Laboratory of Metrology and Testing in Paris.

But the process could be delayed until the 2018 meeting of the BIPM general conference. All four units are intertwined, so the plan is to redefine them all at once, and the kilogram is causing problems. There are two rival approaches to its redefinition: a watt balance, which would balance a test mass against Earth's gravity in terms of electrical power, and a precise count of atoms in a sphere of silicon. The two approaches give slightly different answers. Piquemal says that the differing approaches need to be reconciled before the units can be redefined. ■

1. Fricke, L. *et al*. Preprint at http://arxiv.org/abs/1312.5669 (2013).
2. Giblin, S. P. *et al. Nature Commun.* **3,** 930 (2012).
3. Pekola, J. P. *et al. Rev. Mod. Phys.* **85,** 1421–1472 (2013).



This remnant of Tycho's supernova was created by a type Ia supernova in 1572.

X-RAY: NASA/CXC/SAO; INFRARED: NASA/JPL-CALTECH; OPTICAL: MPIA/CALAR ALTO/O. KRAUSE *ET AL.*

ASTROPHYSICS

Kepler clue to supernova puzzle

Two white dwarfs favoured as precursors of type Ia supernovae.

BY RON COWEN

They are cosmic detonations that briefly outshine the light of entire galaxies. And they were a crucial tool in the discovery of dark energy, the force that is accelerating the expansion of the Universe. Yet the process that gives rise to type Ia supernovae has remained mysterious.

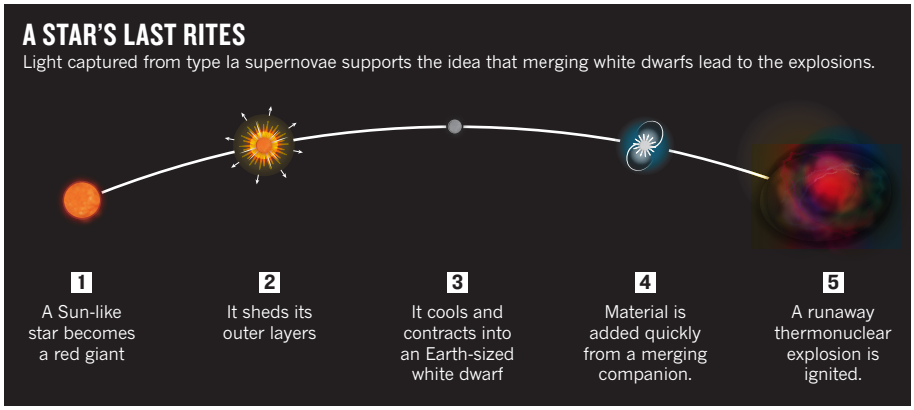
Now, light from two of these stellar explosions has been captured in finer temporal detail than ever before, and the data are adding weight to an emerging view: that the explosions result from the merger of two white dwarfs, the burnt-out, Earth-sized remnants of Sun-like stars. The finding erodes a long-standing view that type Ia supernovae arise from a single white dwarf accruing material from an ordinary companion star, either a Sun-like star or an elderly, bloated red giant.

The data have come from an unlikely source: NASA's Kepler mission, the space telescope that searched for alien planets by staring at some 150,000 stars in nearby reaches of the Milky Way. Distant galaxies also lurk in the telescope's field of view, and its ability to collect data every 30 minutes, along with its sensitivity to tiny changes in brightness, made it ideal for recording the rise and fall of light emitted during supernovae.

Robert Olling, an astronomer at the University of Maryland in College Park, was lucky enough to find two type Ia supernovae after a two-year survey of some 400 galaxies in Kepler's field. He reported them on 8 January at a meeting of the American Astronomical Society near Washington DC. "As a technical tour de force, it's really cool to use Kepler for more than it was intended," says Robert Kirshner, an astronomer at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.

The exceptional smoothness of the Kepler data is helping researchers to distinguish between the two competing explosion scenarios. Both require that a white dwarf takes on material from a companion object until the pressure ignites a runaway thermonuclear explosion. But in the ordinary companion model, the expanding shell of material from the white dwarf would ram into the normal star, generating extra heat and light that would show up as a bump in the first days of a supernova's brightening. No such bump exists in Olling's data.

That essentially rules out all red giant companions, says Olling, because these larger stars would result in a sizeable bump. But the data could still be compatible with smaller, more



Sun-like companions, says Daniel Kasen, an astronomer at the University of California, Berkeley, and a collaborator on the survey. Not only would these stars cause a much smaller bump, but the bump could be missed completely depending on the observer's viewing angle, he says. If the supernova lay between Kepler and the companion star, for example, the associated bump would probably not be seen.

For years, the idea that type Ia supernovae might arise from merging white dwarfs was discounted because the final stages of the merger were thought to occur slowly, over the course of thousands of years. Such a slow accretion of material would be more likely to lead to the formation of a neutron star. Then, from around 2010, simulations began to suggest that the mergers could occur in seconds or minutes, allowing for the sudden pressure change that results in an explosion, says Stan Woosley, a theorist at the University of California, Santa Cruz (see 'A star's last rites').

Craig Wheeler, a supernova theorist at the University of Texas at Austin, says that there are still problems with the merger model. For example, he says, simulations of the mergers often produce highly asymmetric explosions, yet observations so far tend to be more spherical. And spectroscopic observations — which split light into its component wavelengths — have not found as much radiation from ionized iron atoms as merger simulations predict.

Kepler's exoplanet-hunting days ended in May 2013 after mechanical failures prevented it from pointing precisely enough for that task. But Olling says that the craft could continue to hunt for type Ia supernovae because the bright explosions do not require precise pointing.

It will be crucial to make simultaneous observations from ground-based telescopes, he notes, because Kepler only records brightness and cannot split light into spectra. But to perform such joint observations, Kepler will need to point in the opposite direction; Olling hopes that the Kepler team will agree to this. NASA is expected to announce its plans for the impaired spacecraft this summer. ■

CORRECTION

This News story 'Particle-physics papers set free' (*Nature* **505,** 141; 2014) wrongly stated that CERN has decreed that all articles based on its research must be open access. In fact, it is still reviewing its policy. And in the World View 'Academics should not remain silent on hacking' (*Nature* **504,** 333; 2013), the URL for the non-profit organization recruiting experts should have been opencryptoaudit.org.

THE CASE OF THE MISSING HEAT

Sixteen years into the mysterious 'global-warming hiatus', scientists are piecing together an explanation.

BY JEFF TOLLEFSON

The biggest mystery in climate science today may have begun, unbeknownst to anybody at the time, with a subtle weakening of the tropical trade winds blowing across the Pacific Ocean in late 1997. These winds normally push sun-baked water towards Indonesia. When they slackened, the warm water sloshed back towards South America, resulting in a spectacular example of a phenomenon known as El Niño. Average global temperatures hit a record high in 1998 — and then the warming stalled.

For several years, scientists wrote off the stall as noise in the climate system: the natural variations in the atmosphere, oceans and biosphere that drive warm or cool spells around the globe. But the pause has persisted, sparking a minor crisis of confidence in the field. Although there have been jumps and dips, average atmospheric temperatures have risen little since 1998, in seeming defiance of projections of climate models and the ever-increasing emissions of greenhouse gases. Climate sceptics have seized on the temperature trends as evidence that global warming

has ground to a halt. Climate scientists, meanwhile, know that heat must still be building up somewhere in the climate system, but they have struggled to explain where it is going, if not into the atmosphere. Some have begun to wonder whether there is something amiss in their models.

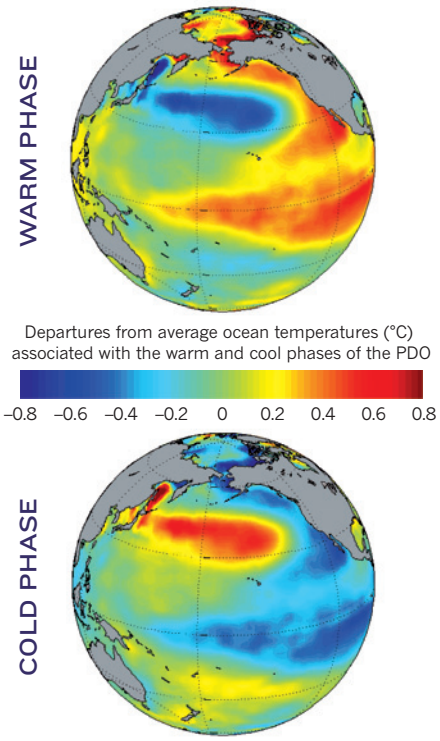
Now, as the global-warming hiatus enters its sixteenth year, scientists are at last making headway in the case of the missing heat. Some have pointed to the Sun, volcanoes and even pollution from China as potential culprits, but recent studies suggest that the oceans are key to explaining the anomaly. The latest suspect is the El Niño of 1997–98, which pumped prodigious quantities of heat out of the oceans and into the atmosphere — perhaps enough to tip the equatorial Pacific into a prolonged cold state that has suppressed global temperatures ever since.

"The 1997 to '98 El Niño event was a trigger for the changes in the Pacific, and I think that's very probably the beginning of the hiatus," says Kevin Trenberth, a climate scientist at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. According to this theory, the tropical Pacific should snap out of its prolonged cold spell in the coming years. "Eventually," Trenberth says, "it will switch back in the other direction."

NATE MANTUA/NOAA

THE FICKLE OCEAN

Every few decades, the Pacific flips between warmer and cooler conditions in the eastern part of the basin and along much of the equator. This cycle is called the Pacific Decadal Oscillation (PDO).



STARK CONTRAST

On a chart of global atmospheric temperatures, the hiatus stands in stark contrast to the rapid warming of the two decades that preceded it. Simulations conducted in advance of the 2013–14 assessment from the Intergovernmental Panel on Climate Change (IPCC) suggest that the warming should have continued at an average rate of 0.21 °C per decade from 1998 to 2012. Instead, the observed warming during that period was just 0.04 °C per decade, as measured by the UK Met Office in Exeter and the Climatic Research Unit at the University of East Anglia in Norwich, UK.

The simplest explanation for both the hiatus and the discrepancy in the models is natural variability. Much like the swings between warm and cold in day-to-day weather, chaotic climate fluctuations can knock global temperatures up or down from year to year and decade to decade. Records of past climate show some long-lasting global heatwaves and cold snaps, and climate models suggest that either of these can occur as the world warms under the influence of greenhouse gases.

But none of the climate simulations carried out for the IPCC produced this particular hiatus at this particular time. That has led sceptics — and some scientists — to the controversial conclusion that the models might be overestimating the effect of greenhouse gases, and that future warming might not be as strong as is feared. Others say that this conclusion goes against the long-term temperature trends, as well as palaeoclimate data that are used to extend the temperature record far into the past. And many researchers caution against evaluating models on the basis of a relatively short-term blip in the climate. "If you are interested in global climate change, your main focus ought to be on timescales of 50 to 100 years," says Susan Solomon, a climate scientist at the Massachusetts Institute of Technology in Cambridge.

But even those scientists who remain confident in the underlying models acknowledge that there is increasing pressure to work out just what is happening today. "A few years ago you saw the hiatus, but it could be dismissed because it was well within the noise," says Gabriel Vecchi, a climate scientist at the US National Oceanic and Atmospheric Administration's Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey. "Now it's something to explain."

TIM GRAHAM/ROBERT HARDING PICTURE LIBRARY

Researchers have followed various leads in recent years, focusing mainly on a trio of factors: the Sun[1], atmospheric aerosol particles[2] and the oceans[3]. The output of energy from the Sun tends to wax and wane on an 11-year cycle, but the Sun entered a prolonged lull around the turn of the millennium. The natural 11-year cycle is currently approaching its peak, but thus far it has been the weakest solar maximum in a century. This could help to explain both the hiatus and the discrepancy in the model simulations, which include a higher solar output than Earth has experienced since 2000.

An unexpected increase in the number of stratospheric aerosol particles could be another factor keeping Earth cooler than predicted. These particles reflect sunlight back into space, and scientists suspect that small volcanoes — and perhaps even industrialization in China — could have pumped extra aerosols into the stratosphere during the past 16 years, depressing global temperatures.

Some have argued that these two factors could be primary drivers of the hiatus, but studies published in the past few years suggest that their effects are likely to be relatively small[4,5]. Trenberth, for example, analysed their impacts on the basis of satellite measurements of energy entering and exiting the planet, and estimated that aerosols and solar activity account for just 20% of the hiatus. That leaves the bulk of the hiatus to the oceans, which serve as giant sponges for heat. And here, the spotlight falls on the equatorial Pacific.

BLOWING HOT AND COLD

Just before the hiatus took hold, that region had turned unusually warm during the El Niño of 1997–98, which fuelled extreme weather across the planet, from floods in Chile and California to droughts and wildfires in Mexico and Indonesia. But it ended just as quickly as it had begun, and by late 1998 cold waters — a mark of El Niño's sister effect, La Niña — had returned to the eastern equatorial Pacific with a vengeance. More importantly, the entire eastern Pacific flipped into a cool state that has continued more or less to this day.

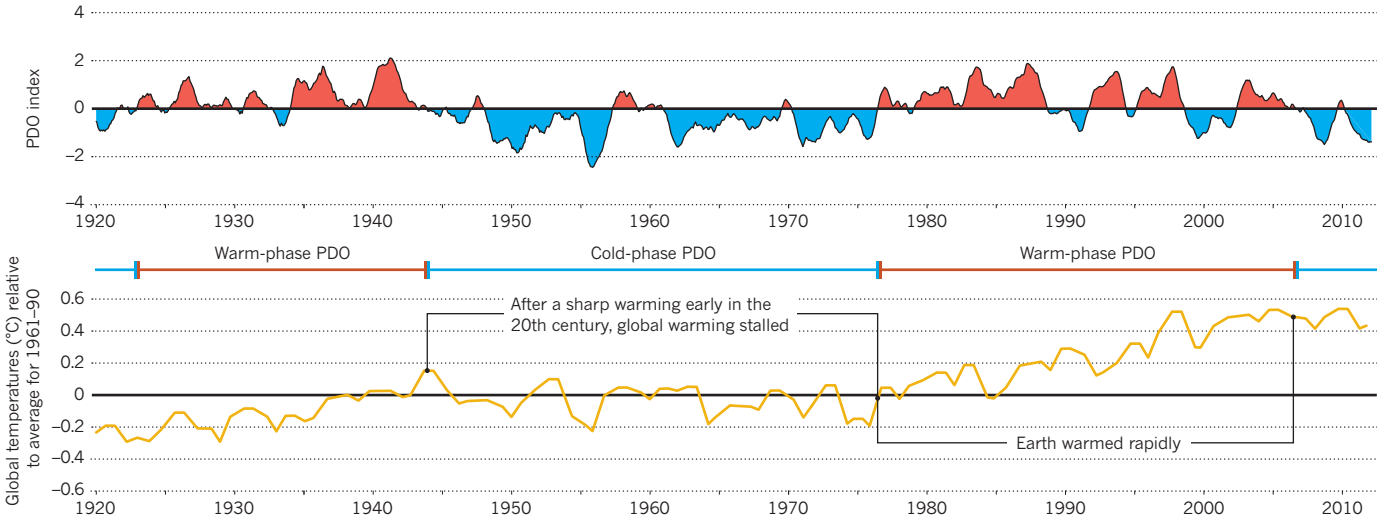
This variation in ocean temperature, known as the Pacific Decadal Oscillation (PDO), may be a crucial piece of the hiatus puzzle. The cycle reverses every 15–30 years, and in its positive phase, the oscillation favours El Niño, which tends to warm the atmosphere (see 'The fickle ocean'). After a couple of decades of releasing heat from the eastern and central Pacific, the region cools and enters the negative phase of the PDO. This state tends towards La Niña, which brings cool waters up from the depths along the Equator and tends to cool the planet. Researchers identified the PDO pattern in 1997, but have only recently begun to understand how it fits in with broader ocean-circulation patterns and how it may help to explain the hiatus.

One important finding came in 2011, when a team of researchers at NCAR led by Gerald Meehl reported that inserting a PDO pattern into global climate models causes decade-scale breaks in global warming[3]. Ocean-temperature data from the recent hiatus reveal why: in a subsequent study, the NCAR researchers showed that more heat moved into the deep ocean after 1998, which helped to prevent the atmosphere from warming[6]. In a third paper, the group used computer models to document the flip side of the process: when the PDO switches to its positive phase, it heats up the surface ocean and atmosphere, helping to drive decades of rapid warming[7].

A key breakthrough came last year from Shang-Ping Xie and Yu Kosaka at the Scripps Institution of Oceanography in La Jolla,

THE PACIFIC'S GLOBAL REACH

As researchers have investigated why global temperatures have not risen much since 1998, many have focused on an ocean cycle known as the Pacific Decadal Oscillation (PDO). During periods when the PDO index is positive and the eastern Pacific is warm, global temperatures have risen quickly. During spells when the PDO index is negative, the warming has stagnated.



UNIV. WASHINGTON/IPCC

California. The duo took a different tack, by programming a model with actual sea surface temperatures from recent decades in the eastern equatorial Pacific, and then seeing what happened to the rest of the globe[8]. Their model not only recreated the hiatus in global temperatures, but also reproduced some of the seasonal and regional climate trends that have marked the hiatus, including warming in many areas and cooler northern winters.

"It was actually a revelation for me when I saw that paper," says John Fyfe, a climate modeller at the Canadian Centre for Climate Modelling and Analysis in Victoria. But it did not, he adds, explain everything. "What it skirted was the question of what is driving the tropical cooling."

That was investigated by Trenberth and John Fasullo, also at NCAR, who brought in winds and ocean data to explain how the pattern emerges[4]. Their study documents how tropical trade winds associated with La Niña conditions help to drive warm water westward and, ultimately, deep into the ocean, while promoting the upwelling of cool waters along the eastern equatorial region. In extreme cases, such as the La Niña of 1998, this may be able to push the ocean into a cool phase of the PDO. An analysis of historical data buttressed these conclusions, showing that the cool phase of the PDO coincided with a few decades of cooler temperatures after the Second World War (see 'The Pacific's global reach'), and that the warm phase lined up with the sharp spike seen in global temperatures between 1976 and 1998 (ref. 4).

"I believe the evidence is pretty clear," says Mark Cane, a climatologist at Columbia University in New York. "It's not about aerosols or stratospheric water vapour; it's about having had a decade of cooler temperatures in the eastern equatorial Pacific."

HEATED DEBATE

Cane was the first to predict the current cooling in the Pacific, although the implications weren't clear at the time. In 2004, he and his colleagues found that a simple regional climate model predicted a warm shift in the Pacific that began around 1976, when global temperatures began to rise sharply[9]. Almost as an afterthought, they concluded their paper with a simple forecast: "For what it is worth the model predicts that the 1998 El Niño ended the post-1976 tropical Pacific warm period."

It is an eerily accurate result, but the work remains hotly contested, in part because it is based on a partial climate model that focuses on the equatorial Pacific alone. Cane further maintains that the trend over the past century has been towards warmer temperatures in the western Pacific relative to those in the east. That opens the door, he says, to the possibility that warming from greenhouse gases is driving La Niña-like conditions and could continue to do so in the future, helping to suppress global warming. "If all of that is true, it's a negative feedback, and if we don't capture it in our models they will overstate the warming," he says.

There are two potential holes in his assessment. First, the historical ocean-temperature data are notoriously imprecise, leading many researchers to dispute Cane's assertion that the equatorial Pacific shifted towards a more La Niña-like state during the past century[10]. Second, many researchers have found the opposite pattern in simulations with full climate models, which factor in the suite of atmospheric and oceanic interactions beyond the equatorial Pacific. These tend to reveal a trend towards more El Niño-like conditions as a result of global warming. The difference seems to lie, in part, in how warming influences evaporation in areas of the Pacific, according to Trenberth. He says the models suggest that global warming has a greater impact on temperatures in the relatively cool east, because the increase in evaporation adds water vapour to the atmosphere there and enhances atmospheric warming; this effect is weaker in the warmer western Pacific, where the air is already saturated with moisture.

Scientists may get to test their theories soon enough. At present, strong tropical trade winds are pushing ever more warm water westward towards Indonesia, fuelling storms such as November's Typhoon Haiyan, and nudging up sea levels in the western Pacific; they are now roughly 20 centimetres higher than those in the eastern Pacific. Sooner or later, the trend will inevitably reverse. "You can't keep piling up warm water in the western Pacific," Trenberth says. "At some point, the water will get so high that it just sloshes back." And when that happens, if scientists are on the right track, the missing heat will reappear and temperatures will spike once again. ■ SEE EDITORIAL P.261

Jeff Tollefson *covers climate, energy and the environment for* Nature.

1. Lean, J. L. & Rind, D. H. *Geophys. Res. Lett.* **36,** L15708 (2009).
2. Hansen, J., Sato, M., Kharecha, P. & von Schuckmann, K. *Atmos. Chem. Phys.* **11,** 13421–13449 (2011).
3. Meehl, G. A., Arblaster, J. M., Fasullo, J. T., Hu, A. & Trenberth, K. E. *Nature Clim. Change* **1,** 360–364 (2011).
4. Trenberth, K. E. & Fasullo, J. T. *Earth's Future* http://dx.doi.org/10.1002/2013EF000165 (2013).
5. Feulner, G. & Rahmstorf, S. *Geophys. Res. Lett.* **37,** L05707 (2010).
6. Balmaseda, M. A., Trenberth, K. E. & Källén, E. *Geophys. Res. Lett.* **40,** 1754–1759 (2013).
7. Meehl, G. A., Hu, A., Arblaster, J. M., Fasullo, J. & Trenberth, K. E. *J. Clim.* **26,** 7298–7310 (2013).
8. Kosaka, Y. & Xie, S.-P. *Nature* **501,** 403–407 (2013).
9. Seager, R. *et al.* in *Earth's Climate: The Ocean-Atmosphere Interaction. Geophys. Monogr. Ser.* **147,** 105–120 (2004).
10. DiNezio, P., Clement, A. & Vecchi, G. A. *Eos* **91,** 141–152 (2010).

STARS: PIXELPARTICLE/SHUTTERSTOCK; ILLUSTRATION: JASIEK KRZYSZTOFIAK/*NATURE*

BY RON COWEN

THE HEART OF DARKNESS

THE SUPERMASSIVE BLACK HOLES THAT LIE AT THE CENTRE OF EVERY LARGE GALAXY ARE FULL OF MYSTERIES. BUT ASTRONOMERS ARE FINALLY GETTING A CLEAR LOOK.

Many of the astronomers and physicists invited to the meeting feared for their safety. Others felt that the event should be cancelled outright. To hold a conference in Dallas, Texas, only weeks after US President John F. Kennedy had been assassinated there — it just seemed disrespectful.

In the end, the first Texas Symposium on Relativistic Astrophysics went ahead as scheduled, starting on 16 December 1963, and most of the invited scientists did go — after the mayor of Dallas sent them a telegram urging their attendance. But the shadow cast by Kennedy's death added to the already surreal mood as they grappled with a phenomenon that seemed unfathomable.

That year, observers had discovered that a collection of mysterious 'quasi-stellar' objects, dubbed quasars, were not just oddball versions of ordinary stars. They were cosmically distant, glowing with radiation that had travelled for billions of years to reach Earth. They were prodigiously bright, able to outshine 100 galaxies containing billions of normal stars. And they were astonishingly small for such bright objects — no bigger than our own Solar System. The presence of so much energy in so small a volume would bend space-time, as described by Albert Einstein's general theory of relativity, and might even cause the matter there to collapse into a gigantic black hole: an exotic possibility that at the time seemed like pure science fiction.

"Quasars really changed everything," says Michael Turner, a cosmologist at the University of Chicago, Illinois, who gave a speech commemorating the 50th anniversary of that inaugural meeting last month at the 27th Texas symposium, again in Dallas. Einstein's theory, which until the 1960s had been considered a niche idea with little to do with practical astronomy, was pushed to the fore. "The floodgates had opened," says Turner: observations soon proved that the Universe was stranger and more violent than astronomers had ever imagined. Explosions and eruptions were commonplace. And Solar System-sized black holes with masses measured in millions or billions of Suns turned out to lie not just inside quasars, but at the centre of every large galaxy in the cosmos — including our own.

THE UNIVERSE WAS STRANGER AND MORE VIOLENT THAN ASTRONOMERS HAD EVER IMAGINED.

As last month's symposium made clear, giant black holes still pose many puzzles, ranging from how they produce and release enormous amounts of energy to how they grew rapidly in the early Universe. Researchers are now starting to glean important clues from instruments including NASA's Nuclear Spectroscopic Telescope Array (NuSTAR), which was launched in mid-2012 as the first spacecraft dedicated to studying these objects. And this year astronomers will get a rare chance to study the eating habits of the black hole at the centre of our own Galaxy, when it feasts on a cloud of gas set to stray too close to its gravitational trap.

The basics of black holes' energy production are now well established (see 'Accretion power'). Stars, gas and dust moving through the core of a galaxy get pulled in and compressed by the black hole's gravity, growing hotter and hotter as they spiral inwards, forming an accretion disk. By the time the superheated material approaches a spinning black hole's event horizon — the point of no return, beyond which even light cannot escape — up to 42% of its mass has been converted to energy.

That energy emerges in the form of heat, light and, often, jets of high-speed particles that rocket in opposite directions perpendicular to the accretion disk. These jets can extend for thousands or even millions of parsecs. If one happens to be aimed directly at Earth, astronomers see the object as a quasar. If the jets point sideways instead, astronomers see the object as a galaxy with a very bright 'active galactic nucleus'. And if the black hole's food supply is somehow restricted, so that it accretes very little gas and dust, the object is effectively invisible.

Within that general picture, however, the details can be perplexing. Starting in 2006, for example, several sky surveys began to indicate that jets were emerging from their parent black holes with three times more energy than was contained in the original fuel, producing what seemed to be a gross violation of the conservation of energy.

MAGNETIC BOOST

At last month's conference, physicist Roger Blandford of Stanford University in California described a possible solution based on simulations of jet formation[1,2]. He and his colleagues imagine a rapidly spinning black hole with a strong magnetic field, properties that are difficult to detect directly but are theoretically plausible. The lines of the magnetic field are assumed to go out to great distances, threading through the accretion disk like stiff wires and dragging the disk's gas along with them as they rotate. The simulations show that under the right circumstances, the magnetic field can transfer enough of the black hole's rotational energy into the disk to power the anomalously strong jets.

NuSTAR recently made the first definitive measurement of a supermassive black hole, revealing that it is spinning very fast indeed. This work was prompted by simulations that suggested a way to gauge the rotation of a black hole using X-rays emitted from near the event horizon. Rapidly spinning black holes should pull material closer to that horizon and subject it to intense gravity that would shift escaping X-rays to redder, less energetic wavelengths.

Although astronomers had seen hints of this gravitational imprint with earlier X-ray telescopes, they could not rule out the possibility that gas clouds were blanketing the accretion disk and confounding the result. But NuSTAR is sensitive to X-rays that have ten times higher energies than its predecessors could measure, and that punch through any such clouds. At the December meeting, NuSTAR chief scientist Fiona Harrison, an astronomer at the California Institute of Technology in Pasadena, reported seeing a clear signal of red-shifted X-rays from a relatively nearby spiral galaxy known as NGC 1365. Taken together with measurements at lower X-ray energies made by the European Space Agency's XMM-Newton satellite, the observations showed that NGC 1365's central black hole was spinning at nearly the maximum rate allowed by Einstein's theory[3]. It had enough rotational energy to tear apart its entire home galaxy, if that energy could somehow be unleashed.

NGC 1365 may not be typical. But as NuSTAR and future spacecraft begin to measure black-hole spins further back in time, Harrison says,

ACCRETION POWER

At the centre of every large galaxy lives a giant black hole that swallows gas or dust clouds that stray too close. As matter spirals inwards, it is compressed into an accretion disk. By the time it falls into the black hole, the matter is so hot that much of its mass is converted to energy, which emerges as heat, light and jets of high-energy particles.

NATURE.COM
Read more about the Milky Way's black hole:
go.nature.com/hm8wr9

the data may shed light on another conundrum. Astronomers have found quasars that are powered by billion-solar-mass black holes dating back to some 750 million years after the Big Bang, when the Universe was less than 6% of its current age. How did they get so big so fast?

A black hole's spin rate may be a kind of fossil trace of its formation, Harrison explains. Supermassive black holes are too big to have been formed by a star collapsing under its own gravity, like stellar-mass black holes. If the giant black holes were built from many smaller ones, each merger would have brought together black holes spinning in random directions. After millions or billions of years of such collisions, the full-grown beast would have a net spin close to zero. But if the giant black hole had been built by the merger of just a few medium-sized objects, the growth could have been quicker, the spins would not necessarily have cancelled one another out, and the net rotation could be quite high.

The near-maximum spin of the black hole in NGC 1365 suggests that at least some supermassive black holes grew through rapid mergers — although that still leaves the question of where the original medium-sized black holes came from.

FAST SPIN, SLOW GROWTH

Yet high spin could be a problem for black-hole growth in the early Universe, says Avi Loeb of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts. A rapidly rotating black hole tends to drag the inner edge of the accretion disk along with it, pulling it inwards, so the infalling matter has to trace out a longer, slower spiral to reach the event horizon than it would if the black hole were spinning slowly. And that provides more time for its mass to be converted into radiation instead of adding to the hole's mass.

It is conceivable that strong magnetic fields came to the rescue, says Loeb. By transferring the black hole's rotational energy to the outer disk, they could quickly slow its spin, allow more matter to dive inwards and help the earliest black holes to pack on mass. If that is so, then future measurements will show that supermassive black holes have relatively modest spins.

But Loeb's favourite model for how black holes could grow in a hurry involves episodes in which the monster gorges itself on a stream of material so dense and opaque that photons do not have enough time to leak out before the gas makes its final plunge. The radiation is carried inwards instead of escaping, and the black hole swallows its energy as extra mass[4].

Sometimes, a strong magnetic field can stunt a black hole, rather than helping it to grow. That could be what is happening to Earth's closest giant black hole, Sagittarius A*, which lies just 8,300 parsecs from Earth at the heart of the Milky Way. As such objects go, our local specimen is on the small side, with a mass of only four million Suns. And its emissions are minimal.

The question is, why? It may simply be that there is not much gas and dust in the Milky Way's centre for the black hole to swallow. Or maybe something else is at work, says Mitchell Begelman, an astrophysicist at the University of Colorado Boulder. "There is a lot of interesting speculation that some accretion flows are 'magnetically arrested'," he says.

Last year, for example, NuSTAR discovered a magnetar — a highly magnetized neutron star — in an orbit close enough to Sagittarius A* for astronomers to use it to probe the black hole's magnetic field. A close examination of the magnetar's radio emissions shows that the magnetic field surrounding Sagittarius A* is both sizable and highly ordered[5] — perhaps enough to block the black hole's food supply and put it on a near-starvation diet.

Our black hole does occasionally get a little nourishment. Observers are hoping to watch what happens this March, when a distended object called G2 is predicted to come dangerously close to Sagittarius A*. The object, either a gas cloud or a star with a distended gaseous envelope, will be torn apart by the black hole's gravitational tidal forces. If it is gas, the resulting fireworks could be spectacular. But if G2 is a star, the chances of fireworks will be slimmer: it will keep a firmer grip on the gas and less material will fall in, says Andrea Ghez, an astronomer at the University of California, Los Angeles (see *Nature* **495,** 296–298; 2013).

AS SUPERMASSIVE BLACK HOLES GO, THE ONE IN THE MILKY WAY IS ON THE SMALL SIDE.

Either way, astronomers should get a better understanding of what really happens when something falls into a giant black hole. And they may well have a preview in the next few months. In observations unveiled at the Texas meeting, NuSTAR showed that the neighbourhood of Sagittarius A* contains an assortment of small, stellar-sized black holes and neutron stars.

"It's a rare treat that we've been given," says astrophysicist Zoltán Haiman of Columbia University in New York City, who has helped to carry out simulations which suggest that G2's fateful journey may lead to a collision with one of the small black holes[6].

Sagittarius A* promises even more excitement as astronomers gain new observational tools. Over the next few years, all 64 of the radio dishes from the Atacama Large Millimeter/submillimeter Array in northern Chile are expected to join other radio telescopes around the world to create an Earth-sized network. This combination could get an ultra-high-resolution snapshot of how the black hole bends radiation from objects on its far side into a thin ring, or shadow, around Sagittarius A*. Everyone expects the shape of the shadow to conform to the predictions of Einstein's theory. But if it doesn't — if general relativity does not correctly describe space-time around a black hole — the network could offer crucial clues about what theory should replace it.

"That's the big-picture question," says Jonathan McKinney, a physicist at the University of Maryland in College Park. Fifty years after the first Texas symposium, "everyone wants to know if Einstein was right". ■

Ron Cowen *is a freelance writer based in Silver Spring, Maryland.*

1. McKinney, J. C., Tchekhovskoy, A. & Blandford, R. D. *Mon. Not. R. Astron. Soc.* **423,** 3083–3117 (2012).
2. Sikora, M. & Begelman, M. C. *Astrophys. J. Lett.* **764,** L24 (2013).
3. Risaliti, G. *et al. Nature* **494,** 449–451 (2013).
4. Wyithe, J. S. B. & Loeb, A. *Mon. Not. R. Astron. Soc.* **425,** 2892–2902 (2012).
5. Eatough, R. P. *et al. Nature* **501,** 391–394 (2013).
6. Bartos, I., Haiman, Z., Kocsis, B. & Márka, S. *Phys. Rev. Lett.* **110,** 221102 (2013).

COMMENT

REGULATION Data suggest the FDA is overcautious on consumer genomics **p.286**

DEVELOPMENT Why policy-makers must admit that water is finite **p.288**

ASTRONOMY Planetarium show puts dark Universe at the centre of the action **p.290**

FUNDING Grant applications should feature multimedia presentations **p.291**

ILLUSTRATION BY PETE ELLIS/DRAWGOOD.COM

Time to leave GDP behind

Gross domestic product is a misleading measure of national success. Countries should act now to embrace new metrics, urge **Robert Costanza** and colleagues.

Robert F. Kennedy once said that a country's gross domestic product (GDP) measures "everything except that which makes life worthwhile". The metric was developed in the 1930s and 1940s amid the upheaval of the Great Depression and global war. Even before the United Nations began requiring countries to collect data to report national GDP, Simon Kuznets, the metric's chief architect, had warned against equating its growth with well-being.

GDP measures mainly market transactions. It ignores social costs, environmental impacts and income inequality. If a business used GDP-style accounting, it would aim to maximize gross revenue — even at the expense of profitability, efficiency, sustainability or flexibility. That is hardly smart or sustainable (think Enron). Yet since the end of the Second World War, promoting GDP growth has remained the primary national policy goal in almost every country[1].

Meanwhile, researchers have become much better at measuring what actually does make life worthwhile. The environmental and social effects of GDP growth can be estimated, as can the effects of income inequality[2]. The psychology of human well-being can now be surveyed comprehensively and quantitatively[3,4]. A plethora of experiments has produced alternative measures of progress (see Supplementary Information; go.nature.com/bnquxn).

The chance to dethrone GDP is now in sight. By 2015, the UN is scheduled to announce the Sustainable Development Goals, a set of international objectives to improve global well-being. Developing integrated measures of progress attached to these goals offers the global community the opportunity to define what ▶

NATURE.COM
For more on sustainable development goals:
go.nature.com/ttay1n

sustainable well-being means, how to measure it and how to achieve it. Missing this opportunity would condone growing inequality and the continued destruction of the natural capital on which all life on the planet depends.

DETHRONING GDP

When GDP was instituted seven decades ago, it was a relevant signpost of progress: increased economic activity was credited with providing employment, income and amenities to reduce social conflict and prevent another world war.

But the world today is very different from the one faced by the global leaders who met to plan the post-war economy in 1944 in Bretton Woods, New Hampshire. The emphasis on GDP in developed countries now fuels social and environmental instability. It also blinds developing countries to possibilities for more-sustainable models of development.

Soaring economic activity has depleted natural resources. Much of the generated wealth has been unequally distributed, leading to a host of social problems[5]. The philosopher John Stuart Mill noted more than 200 years ago that, once decent living standards were assured, human efforts should be directed to the pursuit of social and moral progress and the increase of leisure, not the competitive struggle for material wealth. Or as the economist John Kenneth Galbraith once observed: "To furnish a barren room is one thing. To continue to crowd in furniture until the foundation buckles is quite another."

The limits of GDP are now clear. Increased crime rates do not raise living standards, but they can lift GDP by raising expenditures on security systems. Despite the destruction wrought by the Deepwater Horizon oil spill in 2010 and Hurricane Sandy in 2012, both events boosted US GDP because they stimulated rebuilding.

WEIGHING THE ALTERNATIVES

Alternative measures of progress can be divided into three broad groups (see Supplementary Information). Those in the first group adjust economic measures to reflect social and environmental factors. The second group consists of subjective measures of well-being drawn from surveys. The third group relies on weighted composite indicators of well-being including housing, life expectancy, leisure time and democratic engagement.

Adjusted economic measures. These are expressed in monetary units, making them more readily comparable to GDP. Such indices consider annual income, net savings and wealth. Environmental costs and benefits (such as destroying wetlands or replenishing water resources) can also be factored in. One example is the genuine progress indicator (GPI). This metric is calculated by starting with personal consumption expenditures, a measure of all spending by individuals and a major component of GDP, and making more than 20 additions and subtractions to account for factors such as the value of volunteer work and the costs of divorce, crime and pollution[6].

Crucially, unlike other measures in the first group, GPI considers income distribution. A dollar's worth of increased income to a poor person boosts welfare more than a dollar's worth of increased income does for a rich person. And a big gap between the richest and the poorest in a country — as in the United States and, increasingly, in China and India — correlates with social problems, including higher rates of drug abuse, incarceration and mistrust, and poorer physical and mental health[5].

"GDP is dangerously inadequate as a measure of quality of life."

These adjustments matter. A 2013 study[2] comparing the GDP per capita and the GPI per capita of 17 countries comprising just over half the global population found startling divergences between the two metrics. The measures were highly correlated from 1950 until about 1978, when they moved apart as environmental and social costs began to outweigh the benefits of increased GDP (see 'Genuine progress flattens'). Tellingly, life satisfaction is highly correlated with GPI per capita, but not with GDP per capita.

Some governments are taking this seriously. Two US states, Vermont and Maryland, have in the past three years adopted GPI as a measure of progress and have implemented policies specifically aimed at improving it.

Subjective measures of well-being. The most comprehensive of these is the World Values Survey (WVS), which covers about 70 countries and includes questions about how satisfied people are with their lives. Starting in 1981, the WVS is conducted in 'waves', the sixth of which is currently in progress. Another example is the gross national happiness index used in Bhutan. This measure uses elaborate surveys that ask how content people feel in nine domains: psychological well-being, standard of living, governance, health, education, community vitality, cultural diversity, time use and ecological diversity.

Subjective well-being has been highly studied, and has even been recommended as the most appropriate measure of societal progress[7]. But subjective indicators are tricky to compare across societies and cultures. For example, self-reported health tracks with clinically reported rates of morbidity and mortality within countries but not across them[8]. And people are not always aware of the things that contribute to their well-being. Few of us give credit to ecosystem services for water supply and storm protection, for example.

Weighted composite measures of several indicators. A comprehensive picture of sustainable societal well-being should integrate subjective and objective indicators[9] (see Supplementary Information, Figure S1), as these measures begin to do. One example is the Happy Planet Index, introduced by the New Economics Foundation in 2006. This multiplies life satisfaction by life expectancy and divides the product by a measure of ecological impact.

Other indices in the third group combine a range of variables, such as income, housing, jobs, health, civic engagement, safety and life satisfaction. The Better Life Index, developed by the Organisation for Economic Co-operation and Development, maintains a website that allows users to choose how to weight variables, revealing how the emphasis on different variables can influence countries' rankings.

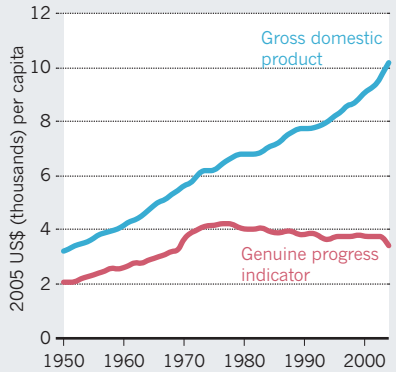
Many other experiments are under way (see www.wikiprogress.org). None of these measures is perfect, but collectively they offer the building blocks for something much better than GDP.

WHY ARE WE STUCK?

There is broad agreement that global society should strive for a high quality of life that is equitably shared and sustainable. Several groups and reports have concluded that GDP is dangerously inadequate as a measure of quality of life — including those published by the French government's 2008 Commission on the Measurement of Economic Performance and Social Progress[10], the Frederick S. Pardee Center for the Study of the Longer-Range Future[11] and the European Commission's ongoing Beyond GDP initiative. That conclusion was also echoed in 'The Future We Want', the declaration

GENUINE PROGRESS FLATTENS

World GDP has soared since 1950, but a metric for life satisfaction called GPI has not.



SOURCE: REF. 2

LEFT: IZZET KERIBAR/IML IMAGE GROUP/EYEVINE; RIGHT: ABE FOX/AP

Bhutan has measured citizens' well-being using gross national happiness since 2008 (left); GDP has been in use since the 1944 Bretton Woods meeting (right).

of the 2012 Rio+20 UN Conference on Sustainable Development agreed to by all UN member states.

Nonetheless, GDP remains entrenched[1]. Vested interests are partly responsible. Former US President Bill Clinton's small move towards a 'green GDP', which factored in some of the environmental consequences of growth, was killed by the coal industry. However, much of the problem is that no alternative measure stands out as a clear successor.

Creating that successor will require a sustained, transdisciplinary effort to integrate metrics and build consensus. One potential vehicle for doing this is the setting up of the UN Sustainable Development Goals (SDGs), a process that is now under way to replace the Millennium Development Goals (MDGs). Established in 2000, the MDGs comprise eight basic targets that include eradicating extreme poverty and establishing universal primary education, gender equality and environmental sustainability. Currently both the MDGs and the suggested SDGs are only lists of goals with isolated indicators. But the SDG process can and should be expanded to include comprehensive and integrated measures of sustainable well-being[12].

If undertaken with sufficiently broad participation, the hunt for the successor to GDP might be completed by 2015. There are significant barriers to doing this, including bureaucratic inertia and the tendency of governments, academia and other groups to work in isolation. These barriers can be overcome with dedicated leadership. Crucially, people can now communicate across the globe with an ease unthinkable in the days of Bretton Woods.

Any 'top-down' process must be supplemented with a 'bottom-up' engagement of civil society that includes city and regional governments, non-governmental organizations, business and other parties. We recently formed the Alliance for Sustainability and Prosperity (www.asap4all.com) to do just that. This web-based 'network of networks' can communicate research about sustainable quality of life and the elements that contribute to it (see Supplementary Information), and so help to build consensus among the thousands of groups now concerned with these issues.

The successor to GDP should be a new set of metrics that integrates current knowledge of how ecology, economics, psychology and sociology collectively contribute to establishing and measuring sustainable well-being. The new metrics must garner broad support from stakeholders in the coming conclaves.

It is often said that what you measure is what you get. Building the future we desire requires that we measure what we want, remembering that it is better to be approximately right than precisely wrong. ■

Robert Costanza *and* **Ida Kubiszewski** *are at the Crawford School of Public Policy, Australian National University, Canberra.* **Enrico Giovannini** *is in the Department of Economics and Finance, University of Rome Tor Vergata, and minister of labour and social policies in the Italian government.* **Hunter Lovins** *is at Natural Capital Solutions, Longmont, Colorado.* **Jacqueline McGlade** *is at University College London, and the United Nations Environment Program, Nairobi, Kenya.* **Kate E. Pickett** *is in the Department of Health Sciences, University of York, UK.* **Kristín Vala Ragnarsdóttir** *is at the Institutes of Earth Sciences and Sustainable Development Studies, University of Iceland, Reykjavík.* **Debra Roberts** *is in the Environmental Planning and Climate Protection Department, eThekwini Municipality, Durban, South Africa.* **Roberto De Vogli** *is at the University of California, Davis.* **Richard Wilkinson** *is in the Division of Epidemiology and Public Health, University of Nottingham, UK.*
e-mail: robert.costanza@anu.edu.au

1. Van den Bergh, J. C. J. M. *J. Econ. Psychol.* **30,** 117–135 (2009).
2. Kubiszewski, I. *et al. Ecol. Econ.* **93,** 57–68 (2013).
3. Diener, E. & Suh, E. M. in *Well-Being: The Foundations of Hedonic Psychology* (eds Kahneman, D., Diener, E. & Schwarz, N.) 434–450 (Russell Sage Foundation, 2003).
4. Seligman, M. E. P. *Flourish: A Visionary New Understanding of Happiness and Well-being* (Atria, 2012).
5. Wilkinson, R. G. & Pickett, K. *The Spirit Level: Why Greater Equality Makes Societies Stronger* (Bloomsbury, 2009).
6. Talberth, J., Cobb, C. & Slattery, N. *The Genuine Progress Indicator 2006: A Tool for Sustainable Development* (Redefining Progress, 2007).
7. Layard, R. *Happiness: Lessons from a New Science* (Penguin, 2005).
8. Barford, A., Dorling, D. & Pickett, K. *Social Sci. Med.* **70,** 496–497 (2010).
9. Costanza, R. *et al. Ecol. Econ.* **61,** 267–276 (2007).
10. Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. *Report by the Commission on the Measurement of Economic Performance and Social Progress* Vol. 12 (Commission on the Measurement of Economic Performance and Social Progress, 2009).
11. Costanza, R., Hart, M., Posner, S. & Talberth, J. *Beyond GDP: The Need for New Measures of Progress* (Boston University, 2009).
12. Griggs, D. *et al. Nature* **495,** 305–307 (2013).

SANDY HUFFAKER/BLOOMBERG/GETTY

A lab technician scans a consumer gene-testing kit.

The FDA is overcautious on consumer genomics

A US drug-agency clampdown is unwarranted without evidence of harm, say **Robert C. Green** and **Nita A. Farahany**.

Since 2007, people have been able to spit in a tube, ship it to a company and later log on to a website to learn what their DNA reveals about non-medical traits such as hair texture and ancestry, as well as whether they carry DNA variants associated with increased risks for diseases including type 2 diabetes and Alzheimer's. More than 500,000 people have bought such tests.

On 22 November last year, the company that has performed the bulk of these tests received a letter from the US Food and Drug Administration (FDA). The FDA ordered 23andMe, based in Mountain View, California, to "immediately discontinue marketing" its testing kit and personal genome services, which the agency says offer medical advice and so require regulatory approval. "Serious concerns are raised if test results are not adequately understood by patients or if incorrect test results are reported," the FDA wrote. Notably, it said, genetic results could drive consumers to take extreme steps, such as having unnecessary surgery to prevent cancer. Consumers might also abandon or alter prescribed treatments without consulting health professionals, the letter alleged.

Two weeks after receiving the letter, 23andMe took steps to allay the FDA's concerns. It continues to offer DNA testing and analysis, but no longer provides new consumers with genetic interpretations that relate to health.

The FDA invoked the precautionary principle — acting on the basis of speculations of potential harm rather than reported harm. Although the FDA requires manufacturers of drugs and diagnostic devices to prove safety and efficacy before marketing, we feel that this approach is unwarranted in regulating 23andMe's personal genome service.

In its earlier warning letters, the agency said that genomic health reporting "appears to meet the definition" of a medical device. In its November correspondence, it states that 23andMe's service is a medical device. The FDA now claims jurisdiction over companies that provide health-related interpretations with genomic data (but not those that provide genomic data alone). But such interpretations, particularly about common genetic variants, relate only indirectly to preventing or diagnosing disease. In this sense, it is like the inferences drawn about rapid weight loss measured by a bathroom scale.

The FDA cannot reasonably regulate all such indirect information as medical devices. Moreover, as the court cases of *Sorrell* v. *IMS Health* (2011) and *United States* v. *Caronia* (2012) demonstrate, doing so could put FDA regulations in greater tension with the First Amendment of the US Constitution, which protects the rights of individuals to receive information, and of 'commercial speech'[1]. Given this backdrop, the agency should avoid restricting consumer genomic testing unless faced with empirical evidence of harm.

Certainly, there are legitimate concerns about 23andMe's approach. Although the accuracy of the technology used is considered to be high, there are no agreed standards to which the company can conform for validating hundreds of simultaneous variant calls. There is also controversy about how to evaluate the accuracy of risks estimated using multiple variants or across ethnicities[2,3]. And consumers might not read or fully understand the company's clear statements that its tests identify only the most common genetic variants and cannot substitute for genetic testing ordered by physicians to assess specific indications, such as a family history of cancer.

Nonetheless, as scholars who study how individuals respond to their own genetic information, we contend that the FDA's precautionary approach may pose a greater threat to consumer health than the harms that it seeks to prevent. Data from more than 5,000 participants suggest that consumer genomics does not provoke distress or inappropriate treatment.

EMERGING EVIDENCE

Over the past five years, we and others have surveyed people who have received consumer genomics results, asking whether they understood them and whether the findings provoked distress, prompted a visit to a doctor or triggered a change in medication or lifestyle.

In 2009, the Scripps Genomic Health Initiative (SGHI) recruited more than 3,000 individuals from health and technology companies and offered subsidized testing through Navigenics (a company that is now owned by Life Technologies and no longer offers consumer testing). Surveys were administered before the participants received results, and 3 and 12 months afterwards. Responses from more than 2,000 participants showed no measurable changes in anxiety or psychological health[4,5].

Consumers who accept subsidized testing might be less anxious than those who seek it out, but we found similar responses in the Impact of Personal Genomics (PGen) Study funded by the US National Institutes of Health (NIH) and jointly led by R.C.G. and health-behaviour researcher Scott Roberts. In 2012–13, we surveyed more than 1,800 customers from two consumer genomics companies — Pathway Genomics and 23andMe. (Pathway Genomics has subsequently changed its business model to focus on tests ordered by physicians.) Preliminary data suggest that on average, customers were briefly

less anxious than their baseline after receiving their results, and never showed elevated anxiety or distress over the following year. These findings were consistent with those observed in the REVEAL (Risk Evaluation and Education for Alzheimer's Disease) Study, a series of NIH-funded randomized trials carried out between 2000 and 2013. More than 700 volunteers, most of whose family members had been diagnosed with Alzheimer's disease, underwent genetic tests assessing their own risk of this disease, and roughly 40% learned that they have increased risk. But even this potentially frightening disclosure caused only modest and transient distress[6,7].

A 2010 survey by researchers at Johns Hopkins University in Baltimore, Maryland, of more than 1,000 customers of three companies found that just over one-quarter of people share their results with physicians in the first few months after receiving them. Similar findings were obtained by the SGHI and PGen Study[5,8]. The Hopkins and PGen studies also found that fewer than 1% of customers reported altering any prescription medicines on the basis of their results without first consulting a physician (see 'Taking action').

Just as patients sometimes misunderstand physicians during office visits, customers sometimes misconstrue results provided by consumer genomics companies. In the Hopkins survey, participants were asked to interpret hypothetical results. Between 5% and 9% interpreted straightforward messages incorrectly — stating that results showing increased risk instead indicated decreased or equal risk, or vice versa.

The FDA is particularly concerned about how people might respond to learning that they have *BRCA* gene mutations that increase their risk of breast and ovarian cancers. Early evidence suggests that consumers respond appropriately. In a study carried out by 23andMe, the company sent interview invitations to 136 customers who carried pathogenic mutations[9]. Of the 32 who accepted, 14 men and 11 women had learned for the first time through consumer testing that they carried a high-risk *BRCA* mutation. All of the women had consulted health-care professionals with their results, and all but one (who elected for surveillance and not surgery) had their tests repeated. Their behaviour after learning this information was no different from that of people who discover these mutations through medical channels.

These interviews also revealed that 30 family members of those carrying a pathogenic mutation decided to get tested themselves; 13 of them tested positive for the high-risk mutation and so received potentially life-saving information they might not have obtained otherwise.

Clearly this is a very small study and conducted by an interested party. It provides no information about the majority of *BRCA*-positive customers, who were not interviewed. Nor did the study evaluate whether women with undetected mutations and a strong family history of breast cancer might have been falsely reassured (despite clear company messages that their *BRCA* testing is not comprehensive) and, as a result, mistakenly elected not to pursue testing in a medical context. Obtaining a truly representative sample is extremely difficult because consumers who participate in surveys might differ from those who do not, and each of the surveys suggest that early customers of genomics services are wealthier and better educated than the general population, and more likely to be white.

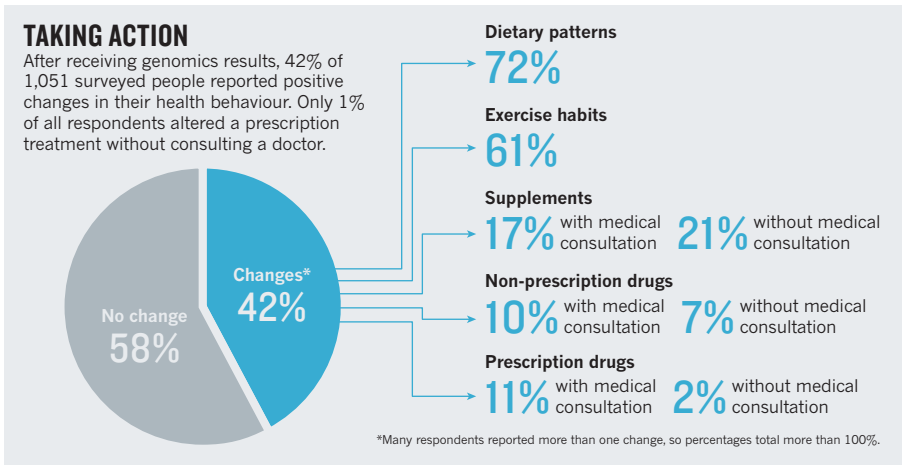
More systematic research is needed to assess the outcomes of consumer genomics testing. The PGen Study results will be available by the end of this year, and further surveys could expand this work to a much larger sample, performing standardized follow-ups on consumers who receive certain high-impact results. However, if consumer genomics is halted, researchers will not be able to continue gathering data to better assess what the benefits and harms could be.

DEMOCRATIZING HEALTH CARE

We find the FDA's precautionary approach to 23andMe particularly troubling because it could presage similar actions against other consumer health products. In its recent guidance on mobile health applications, the FDA left open the possibility that it will regulate as medical devices information-based products such as questionnaires that evaluate the risk of a heart attack or the plethora of fitness trackers that help people to follow their weight, body temperature, heart rate, sleep patterns and more. Many operate as standalone or companion software for predicting risks including the likelihood of sleep disorders, seizures or heart attacks. Downloads and installations of these applications are expected to grow from 156 million in 2012 to 248 million in 2017 (ref. 10).

Such consumer products could democratize health care by enabling individuals to make choices that maximize their own health. They follow the historical trend of patient empowerment that brought informed-consent laws, access to medical records and now direct access to electronic personal health data.

We believe that 23andMe should be more transparent about how accurate its genotyping chips are, and even more forthcoming about the limitations of its computer algorithms used for estimations of risk. But regulatory constraints might stifle consumer genomics and other emerging products that could make society healthier and that do not fit neatly into the model of physician-driven health care. The effects of these products should be monitored but, as long as emerging empirical data show no evidence of harm, we urge the FDA to let consumer genomics testing proceed. ■



Robert C. Green *is in the Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, and the Partners HealthCare Center for Personalized Genetic Medicine, Boston, Massachusetts.* **Nita A. Farahany** *is at the Duke University School of Law and the Duke Institute for Genome Sciences and Policy, Durham, North Carolina.*
e-mails: rcgreen@genetics.med.harvard.edu; farahany@duke.edu

1. Carver, K. H. *Food Drug Law J.* **63,** 151–215 (2008).
2. Kalf, R. R. J. *et al. Genet. Med.* http://dx.doi.org/10.1038/gim.2013.80 (2013).
3. Swan, M. *Genet. Med.* **12,** 279–288 (2010).
4. Bloss, C. S. *et al. Genet. Med.* **12,** 556–566 (2010).
5. Bloss, C. S., Schork, N. J. & Topol, E. J. *N. Engl. J. Med.* **364**, 524–534 (2011).
6. Green, R. C. *et al. N. Engl. J. Med.* **361,** 245–254 (2009).
7. Lautenbach, D. M., Christensen, K. D., Sparks, J. A. & Green, R. C. *Annu. Rev. Genomics Hum. Genet.* **14,** 491–513 (2013).
8. Kaufman, D. J., Bollinger, J. M., Dvoskin, R. L. & Scott, J. A. *J. Genet. Couns.* **21,** 413–422 (2012).
9. Francke, U. *et al. PeerJ* **1,** e8 (2013).
10. Topol, E. *The Creative Destruction of Medicine* (Basic, 2012).

R.C.G. declares competing interests: see go.nature.com/qpucpq for details.

ESPEN RASMUSSEN/PANOS PICTURES

People collect water from a standpipe in Bukavu, Democratic Republic of the Congo.

DEVELOPMENT

Liquid assets

Margaret Catley-Carlson is invigorated by a brace of books on the future of world water supplies.

Water 4.0
DAVID SEDLAK
Yale University Press: 2014.

Blue Future: Protecting Water for People and the Planet Forever
MAUDE BARLOW
The New Press: 2014.

Sweeping, yet very different, perspectives on the human demand for and use of fresh water feature in two new publications. David Sedlak's *Water 4.0* explores what has and should be done to manage water, with efficacy, economics and use-effectiveness as the main metrics. In *Blue Future*, Maude Barlow argues that it is who manages water that is of prime importance; here, the essential metric is the widely cited but not yet well-defined right to water. (I should disclose here that Barlow strongly disapproves of many of the organizations that I am involved with; a photo of me appears in one of her activist videos, I am told.)

Sedlak, co-director of the Berkeley Water Center at the University of California, Berkeley, has contributed a gem to the growing shelf of books on the emerging crises surrounding water, such as the billion people who lack clean supplies. His is an in-depth technical and often political history of water systems with a broad central theme — techniques for water delivery and wastewater treatment, how they work, and what they cost. For example, Sedlak amply covers the water-related infrastructures that are most vulnerable to climate-change-related storms and sea-level rise, such as sewage systems in coastal cities. His focus is on US cities now; he gets there by way of an erudite romp through two millennia of water and sanitation practice and technology.

Sedlak explains that the Roman Empire's aqueduct system ('Water 1.0') delivered different qualities of water for different purposes, using the least clean supplies in latrines and the baths. North Americans today, by contrast, use the same very expensive clean water for all purposes, most of it for watering lawns and flushing toilets. Sedlak quotes Karl Marx's scorn for water management in Victorian England: "they can find no better use for the excrement of four and a half million human beings than to contaminate the Thames with it at heavy expense" (*Capital*, 1867). Marx admired the extensive sewage farms around Paris, which irrigated crops with the effluent — a method still practised round the world. We also see how bad habits developed in the United States: for instance, in 1887 the city of Chicago in Illinois reversed the flow of the Chicago River and sent the sewage to the Mississippi River.

Sedlak is an engineer, but does not overwhelm with technicalities. He marshals chemistry, biology and microbiology to answer numerous pressing questions. For instance, is the nasty film on top of water-filtration systems harmful? (No.) Can we get endocrine disrupters out of water, and stop feminizing male fish? (Perhaps, over time.) Can the natural functions of a very polluted river, such as the Yamuna as it flows through Delhi, be restored through treatment? (Yes, with time, change of habits and investment.)

Sedlak also gives full weight to cultural obstacles such as a reluctance to pay for water. And the economics of the long past, the unsustainable present and the potentially astronomically costly future are clearly put. Among the revelations is the US$13,500 per household it could cost to repair and update drinking-water systems in the United States.

The 'must read' chapter of *Water 4.0* is its last. Here Sedlak explains the book's title in a serious exploration of the decentralized delivery and wastewater options open to industrialized countries and emerging global cities such as Beijing. Sedlak ruefully concedes that a more likely option is even greater investment in today's 'Water 3.0' — centralized, complex, expensive known technology. I wanted him to come out swinging for widespread conversion to greywater systems, in which water from showers and

sinks is recycled to flush toilets. But he is too wise to find a silver bullet in any solution, or to dismiss any out of hand. The undercurrent in this book is that the way forward lies in answers — from the biological to the sociological — that suit the local culture.

Sedlak and Barlow agree that our worldwide failure to value water is at the heart of the problem. Both believe in community participation in decision-making. And both endorse a strong role for public investment in water, particularly in research and in setting guidelines. There, their paths diverge. Where Sedlak seeks to explain the science and technology, Barlow seeks to expose the power relationships.

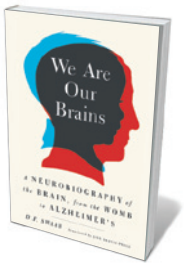
In *Blue Future,* Barlow, a citizens' advocate, makes a passionate plea for the societal change needed to honour the United Nations' 2010 resolution that water and sanitation are a human right. The powerful few, Barlow argues, are blocking this transformation, which would benefit the powerless many. *Blue Future* expands the language of the UN right to encompass all uses of water, and calls for management by a public-trust regime under which all competing uses for watersheds or aquifers would be decided according to a hierarchy of use managed by public agencies.

Barlow has published some 16 books — spirited critiques of issues such as water metering and of bodies such as the World Bank. In *Blue Future* she is also critical of water privatization, which she sees as covering most forms of participation by the private sector, including companies that own no part of the resource or infrastructure but are contracted to carry out government policies. (Only Britain and Chile have privatized water itself — having sold the resource to private companies for onward sale.) And she questions commodification — that is, the use of water markets, price mechanisms, purchase by beverage or mining companies, sale on open markets, and leases to resource extractors, as well as the conversion of utilities to corporatized entities. Looking at one such case in Ireland, Barlow somewhat ingenuously suggests that a water price hike is not needed because Ireland has a lot of water. Yet payments into municipal systems are needed to cover costs: pipes, chemicals, personnel, security and energy for pumping.

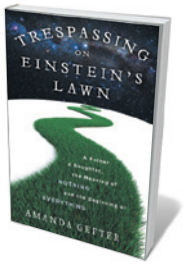
However, Barlow's primary concern (and Sedlak would heartily concur) is that "most political leaders … create policy decisions as if there were no end to water supply". That is the problem in a nutshell. ■

Margaret Catley-Carlson *serves on a dozen boards, advisory committees and juries focused on water and agriculture, including the UN Secretary-General's Advisory Board on Water & Sanitation and the Canadian Water Network.*
e-mail: m.catley-carlson@cgiar.org

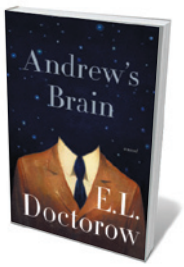
Books in brief



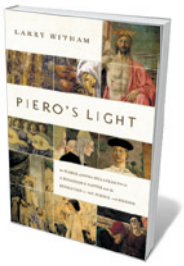
We Are Our Brains: A Neurobiography of the Brain, from the Womb to Alzheimer's
D. F. Swaab (translated by Jane Hedley-Prôle) SPIEGEL & GRAU *(2014)*
In this tour of the human brain's often bumpy terrain, neuroscientist Dick Swaab argues that most of what shapes us happens in the womb. His survey is comprehensive, covering fetal development, sexual differentiation and disorders, birth, early childhood, consciousness, morality, memory and conditions from autism to Alzheimer's disease. The vast scope of this Dutch best-seller demands concision, but Swaab manages to rope each topic and wrestle it to the ground without breaking into a sweat.



Trespassing on Einstein's Lawn: A Father, a Daughter, the Meaning of Nothing and the Beginning of Everything
Amanda Gefter BANTAM *(2014)*
Something can come from nothing. So found Amanda Gefter: a question from her father about the nature of nothing propelled her into science journalism. In this mix of memoir and science, Gefter chronicles her quest to understand the big conundrums through study of the physics literature and meetings with remarkable theoreticians from John Archibald Wheeler to Lisa Randall. Her journey to the insight that reality is in the eye of the beholder is wittily told, but the reverential tone of her starry encounters may jar.

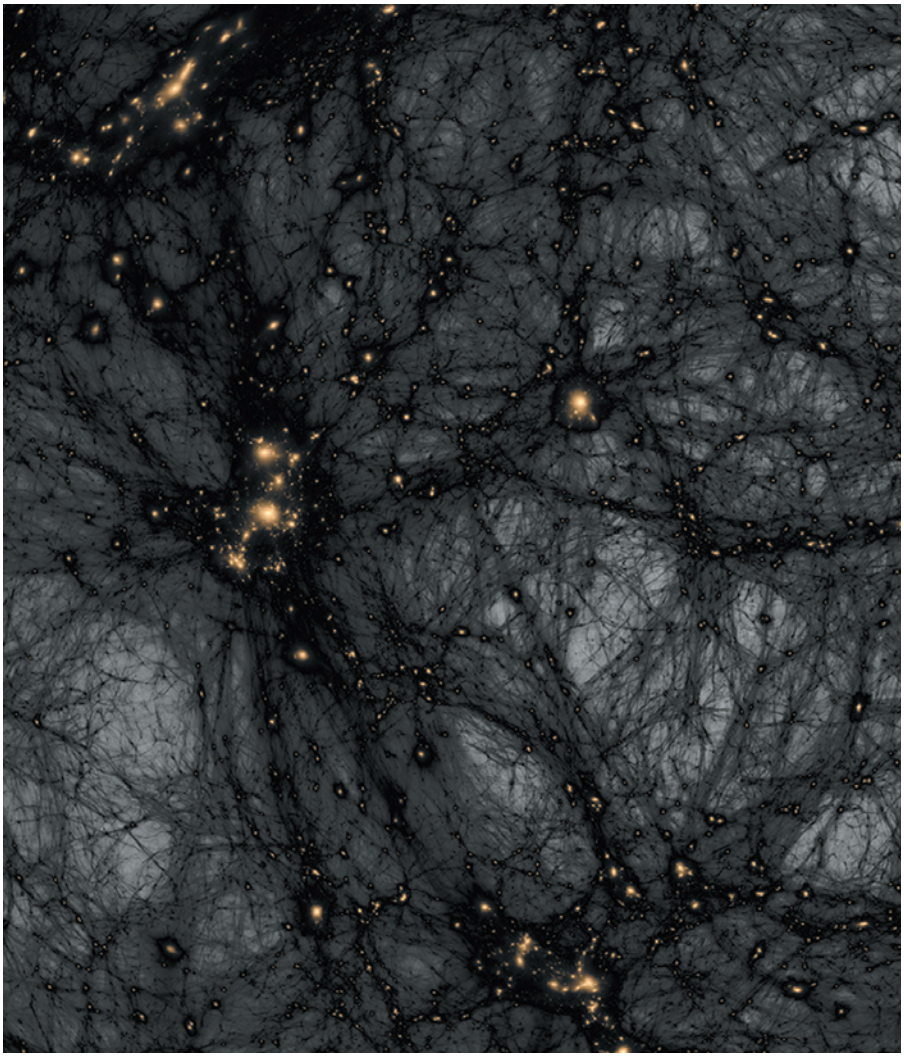


Andrew's Brain: A Novel
E. L. Doctorow RANDOM HOUSE *(2014)*
A cognitive neuroscientist is talking to a psychotherapist — or is it a prison warden? In this spiralling, scientifically savvy narrative on the interplay of brain and mind, distinguished novelist E. L. Doctorow gives us Andrew, an academic recounting his doom-ridden life in snapshots. Doctorow tackles consciousness, free will and memory with elan. The wondrous, sometimes terrifying twists of the human imagination are shot through with gallows humour, thought experiments and even political commentary — and set to a shifting, propulsive rhythm reminiscent of a Philip Glass symphony.



Piero's Light: In Search of Piero della Francesca: A Renaissance Painter and the Revolution in Art, Science and Religion
Larry Witham PEGASUS *(2014)*
The unearthly power of Renaissance artist Piero della Francesca's works is attributable as much to skill in spatial illusion and complex perspective as to artistic brilliance, Larry Witham shows. This study of Piero's impact reveals a mathematician and geometer who helped to bridge the way to Galileo. Along with paintings such as *The Flagellation of Christ* — which fuse maths, classical Platonic science and innovative handling of light — Piero authored influential treatises such the *Little Book on Five Regular Solids*.

Uncharted: Big Data as a Lens on Human Culture
Erez Aiden and Jean-Baptiste Michel RIVERHEAD BOOKS *(2013)*
Thanks to Google Books' digitization of millions of texts dating back centuries, big data is now long data. Erez Aiden and Jean-Baptiste Michel mine the riches using "culturomics", quantifying history by graphing the occurrence of concepts and words in texts over time. In this lively overview, the authors reveal how the robotic historian that they created — the Google Ngram Viewer — has since 2010 been churning out analyses of everything from the efficacy of government censorship to the speed at which society learns. **Barbara Kiser**



Dark Universe **shows dark matter as a web in space, lit up by yellow clusters of galaxies.**

AMNH

ASTRONOMY

The great unseen

Eric Hand views a planetarium show on dark matter and dark energy that is both dislocating and transfixing.

A twinkling, time-lapse whorl that puts you at the centre of the Universe, and a comforting narrative to take you from Earth to outer space in stages — from the known to the unknown. Such is the old-school planetarium experience. But *Dark Universe*, the first new show for four years at the American Museum of Natural History's Hayden Planetarium in New York City, dares to turn the whole thing on its head.

The 23-minute show dislocates the viewer by starting from a disembodied vantage point in the black of outer space. Smudges of light streak by. "Way out here, 10 million light years from planet Earth, every point of light is a galaxy containing billions of stars," intones narrator Neil deGrasse Tyson, the planetarium's director. Zooming in at warp speed, the viewer stumbles on the Milky Way, the Sun and Earth, crashing through the planet's atmosphere to land at the 2.5-metre telescope on Mount Wilson near Los Angeles, California, where Edwin Hubble inferred the expansion of the Universe in the 1920s. The reversed ride is audacious and surprising. But it is the two main focuses of the show — dark matter and dark energy — that really throw you.

Dark Universe
Hayden Planetarium, New York City. Now showing.

These are humbling, even alienating, concepts. Dark matter, the unknown stuff that knits the Universe together, outnumbers the atoms that make up stars, planets and people by a ratio of nearly six to one. Midway through the show, in what he calls its Halloween scene, director Carter Emmart represents dark matter as a spooky black web against a dark-grey background.

Dark energy is even more terrifying. This repulsive force was discovered in the late 1990s, when far-off supernovae were found to be a little more distant than expected. Emmart has fun showing how the supernovae, the known luminosity of which is used to measure distance, explode with strobe-like regularity. Dark energy is accelerating the expansion of the Universe, and threatens one day to hurl other galaxies beyond the view of the Milky Way.

"Lean back in the Hayden's steeply raked seats and let your cosmic insignificance sink in."

I would have liked *Dark Universe* to fast-forward to this depressing but inevitable conclusion, but it does not. The Hayden's shows do not shy away from facts — a previous one depicted the Sun, bloated in old age, encroaching on a scorched Earth — but this particular consequence of dark energy may have been a fact too far. "We don't want to scare our audience," says Emmart. Instead, with streaky vector lines, he tries to explain a difficult, subtle concept: how dark energy's accelerating effect is less apparent on the most distant, oldest galaxies. At the time these galaxies emitted the light that is now reaching Earth, dark energy's repulsion was outmatched by the force of gravity, which was stronger in the dense early Universe.

The rendering of such concepts relies on the digital projectors lining the rim of the planetarium's dome; it would have been impossible with the old-fashioned bulbous star projectors that rose from the centre of the room. Unfettered, Emmart takes full advantage of the third dimension. In a spectacular simulation, we follow the Galileo Probe into the intense pressures of Jupiter's atmosphere. There, in the planet's 'cold-storage locker', astronomers learned about the abundances of primordial hydrogen isotopes that helped to constrain temperatures just after the Big Bang.

Staff at the planetarium hint that its next show, in two or three years, could deal with an altogether cosier subject: the thousands of extrasolar planets discovered in the past two decades, and the astrobiological possibilities therein. More aliens, less alienation. In the meantime, lean back in the Hayden's steeply raked seats and let your cosmic insignificance sink in. The dark Universe, not the blue Earth, is at the centre of it all — and it is man-made telescopes that tell us so. ■

Eric Hand *is* Nature's *US news editor.*

Correspondence

Polar rescue: science was not well served

Following the rescue of passengers from the stranded Russian ship *MV Akademik Shokalskiy*, the expedition's leader, Chris Turney, has spoken out about the importance of science as a driver for the voyage (*Nature* **505,** 133; 2014). No nation hesitates to aid vessels in distress, so why have these events proved so controversial?

Turney said that the "science case" for the voyage was approved by, among others, the Australian Antarctic Division (AAD). But the AAD had no role in assessing, endorsing or approving the scientific merit of the expedition's research plans.

Turney's expedition was a private venture whose research plans were unrelated to Australia's national Antarctic science programme, which is led and managed by the AAD. His voyage aimed to "meld science and adventure" and included as many paying tourists as it did scientists and students (www.spiritofmawson.com).

The rescue disrupted the science and operations programmes of Australia, China and France, who all diverted their ships at the request of the Australian Rescue Coordination Centre, which ran the aid operation. A US icebreaker was also diverted. The financial cost of the exercise is not yet known.

Critics have questioned Turney's claims regarding the scientific importance of the voyage. The trip was relatively brief and seemed to involve the collection of routine samples. By contrast, research projects supported by national polar programmes are multi-year, multinational efforts that rely on sophisticated bespoke equipment.

As a result, Turney's expedition has sparked yet another poorly informed debate on climate science and has seen issues associated with independent Antarctic tourism get conflated with the conduct of Antarctic science. Science is invariably the loser in such cases.

Nick Gales *Australian Antarctic Division, Kingston, Tasmania, Australia.*
nick.gales@aad.gov.au

Shift aims of China's poorer universities

China's universities and research institutes are experiencing the 'Matthew effect', whereby powerful institutions become stronger and weaker ones become even weaker. It is in the country's long-term interest to rectify this by shifting the educational focus of less-prestigious universities to help domestic job seekers.

Chinese academic institutions are clamping down on the recruitment of research staff as the market becomes saturated with domestic PhD graduates. The top-tier universities (known as '985 Project' institutions; see H. Zhang *et al. Res. Policy* **42,** 765–775; 2013) set the most stringent entry requirements, demanding high qualifications from leading and overseas institutions.

The aim seems to be to attract international talent, as outlined in the government's Recruitment Program of Global Experts, with a view to creating several world-class academic universities in China. To this end, the leading universities have already received a large amount of financial and other support. At the same time, 'ordinary' universities are becoming less competitive as they lose the power to draw in resources and the best people.

To restore the balance, we suggest that these less powerful universities should focus on offering vocational education programmes (see Q. Wang *Nature* **499,** 381; 2013) or training students for high-level positions outside academia (see *Nature* **472,** 276–279; 2011) .

Yanhong Tang, Xin Miao *Harbin Institute of Technology, Harbin, China.*
xin.miao@aliyun.com

Rapid progress in producing graphene

Your discussion on developments in graphene research might give the erroneous impression that we are decades away from a commercially viable method of graphene production on an industrial scale (see *Nature* **503,** 327–329; 2013).

In fact, over the past few years the ability to grow large tracts of graphene and transfer it to multiple substrates, as well as doping, patterning and other techniques, have been developed to a point at which industrial production is progressing rapidly. You mention that the company Graphenea produces 15 square metres of graphene per year. But other companies have superior production capabilities: US-based Bluestone Global Tech, for example, has the capacity to produce 20–200 square metres of graphene per day (see go.nature.com/gja2bo).

The production cost you quote of up to US$100,000 per square metre is actually a retail price for just a few square centimetres of material. The cost of producing it is in fact much lower. Production costs have fallen rapidly in the past few months at Bluestone Global Tech, thanks to an increase in production volumes and new transfer techniques.

Kostya S. Novoselov* *University of Manchester, UK.*
kostya@manchester.ac.uk
**On behalf of 4 co-signatories (see go.nature.com/utk5pk for full list).*

Use multimedia in grant applications

The format of grant applications should be updated to incorporate multimedia video. This would help researchers to convey complex topics to grant-review panels.

If time-poor research panels cannot quickly grasp the scientific ideas presented in a paper application, other factors, such as author affiliations and track records, may disproportionately influence project rankings.

We contend that dense proposals presented in a more accessible multimedia format will enhance reviewer comprehension, and so preserve the effectiveness of the merit-based grant-ranking system and boost the research benefit from increasingly scarce public funds.

Researchers already routinely use videos to communicate complex scientific concepts and methods at conferences and in online journals and seminars. Ironically, the US National Institutes of Health even uses YouTube video presentations to instruct applicants on how to prepare paper grant applications.

Michael R. Doran, William B. Lott *Queensland University of Technology, Brisbane, Australia.*
michael.doran@qut.edu.au
Steven E. Doran *University of Illinois, Urbana-Champaign, USA.*

Publish on the basis of quality, not gender

The publication of research papers should be based on quality and merit, so the gender balance of authors is not relevant in the same way as it might be for commissioned writers (see *Nature* **504,** 188; 2013). Neither is the disproportionate number of male reviewers evidence of gender bias.

Having young children may prevent a scientist from spending as much time publishing, applying for grants and advancing their career as some of their colleagues. Because it is usually women who stay at home with their children, journals end up with more male authors on research articles. The effect is exacerbated in fast-moving fields, in which taking even a year out threatens to leave a researcher far behind.

This means that there are likely to be more men in the pool of potential referees.

Lukas Koube *Sherman, Texas, USA.*
lukas.koube@gmail.com

EARTH SCIENCE

River incision revisited

A data-set compilation suggests that measurements of river erosion into rock depend on the observation timescale, casting doubt on whether terraces and other incised landforms faithfully record changes in climate and tectonics. SEE LETTER P.391

ROMAN A. DIBIASE

The pattern and ages of relict landforms, such as terraces, that are incised by rivers provide the best record of changes in Earth's surface elevation over millennial to million-year timescales. As a result, much work has been devoted to determining the age of bedrock landforms[1] and sediment deposited on terraces[2] and in caves[3], to constrain the magnitude and timing of river incision. On page 391 of this issue, Finnegan *et al.*[4] challenge the perception that the ages of relict landforms along incised bedrock rivers (Fig. 1) retain a signature of climatic and tectonic forcing. Instead, the authors argue that the intermittent nature of bedrock river incision means that attempts to measure long-term incision rates are inherently biased by the timescale over which they are averaged.

Bedrock rivers set the pace of landscape evolution and form an important connection between Earth's tectonic and climatic systems. As mountain ranges are uplifted, increases in topographic relief cause rivers to incise bedrock towards a balance between erosion and uplift. Furthermore, the ability of rivers to incise rock and transport sediment is controlled by the size and frequency of floods. Such floods are sensitive to changes in climate that may in turn be driven by tectonic processes. In recent decades, provocative hypotheses[5] regarding the potential coupling of climate, tectonics and erosion have impelled the need for a quantitative understanding of the mechanics of river incision[6].

The idea of a link between the rates of change in surface elevation and the measurement timescale is well known in stratigraphy; this bias is known as the Sadler effect, after its discovery[7] by Earth scientist Peter Sadler. In the sedimentary record, depositional history is not continuous, but is broken by gaps of non-deposition or erosion, owing to the dynamics of sedimentary processes. As a result, measurements of deposition rates over timescales shorter than the longest gap will overestimate long-term averages in a predictable manner[8]. Might a similar bias be present in net erosional landscapes?

Finnegan *et al.* approached this problem by compiling a global data set of incision records from dated terraces, caves and lava flows. They show that, in general, incision rates decrease with increasing measurement interval. The authors combed published studies for field sites that have multiple levels of dated incised landforms spanning a wide range in age, and analysed records from diverse tectonic settings — including the rapidly uplifting middle gorge of the Indus River in the Himalayas and rivers draining the tectonically quiescent Appalachian Mountains in the eastern United States. They found that, in 11 of 14 locations, apparent incision rates increase towards the present day. This relationship holds across four orders of magnitude in landform age, from thousands to tens of millions of years, and is independent of tectonic setting or landform type.

L. C. MALATESTA

Figure 1 | A flight of river terraces in the Tian Shan foreland, China. Finnegan *et al.*[4] argue that river incision rates into bedrock, determined from dated landforms such as these, are biased by the timescale over which they are measured.

The researchers reasoned that the observed scaling relationship is due to the intermittency of river incision, and they employed a random-walk model to show this effect. In the model, as the time window of observation increases, so does the likelihood of experiencing a long period without erosion, and an inverse power law scaling between the inferred incision rate and measurement interval emerges. Finnegan *et al.* propose that the hiatuses in river incision occur when sediment covers and protects the river bed from erosion, and are analogous to gaps in the sedimentary record. This implies that erosional as well as depositional landscapes are subject to an observational bias when deriving rates that average over different time spans.

The similarity in scaling behaviour for sites from a wide range of tectonic settings hints at the existence of a common driver for the perceived acceleration of incision rates, but there are a few factors that complicate this interpretation. First, abandoned flights of terraces and deep valleys that preserve incision markers are the precise landforms expected for landscapes undergoing an actual increase in erosion as a result of changes in climate or tectonics[9]. Further independent information from these landscapes is needed to provide constraints on the plausibility of the steady-state erosion implicitly assumed by Finnegan and colleagues.

Second, although there is little argument that river incision occurs only intermittently,

and that deposition in response to landslides can protect river beds from eroding over millennial timescales[10], it is unclear what physical processes drive hiatuses in incision over timescales of 1 million to 10 million years. Such long timescales span multiple glacial–interglacial climate cycles, and may also reflect a connection to deep Earth processes. Perhaps the most intriguing implication of Finnegan and co-workers' study is the idea that long-term rates of landscape lowering may be more sensitive to the frequency and magnitude of depositional events than to the mechanics of river incision into bedrock.

If the authors' findings hold true, a natural question arises: can changes in the pace of landscape evolution be deciphered from net erosional landscapes? Work on depositional landscapes shows that the preservation bias in one-dimensional stratigraphy disappears when the spatial distribution of both preserved sediments and hiatuses within a basin are taken into account[11–13]. By incorporating a similar spatial averaging in erosional landscapes[12,13], or by accounting for changes in hillslope lowering over time using different chronometers[14], it may be possible to overcome some of these biases.

Finally, to understand the influence of tectonics, land use or climate change on erosion rates, we need a robust way to compare rates measured over different time intervals. Although this is a challenging task, characterizing the degree to which these rates are unsteady, by studying the processes that control erosion and deposition, is essential for interpreting rates measured over different periods of Earth's history and for predicting future change. ■

Roman A. DiBiase *is in the Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.*
e-mail: rdibiase@caltech.edu

1. Leland, J., Reid, M. R., Burbank, D. W., Finkel, R. & Caffee, M. *Earth Planet. Sci. Lett.* **154,** 93–107 (1998).
2. Anderson, R. S., Repka, J. L. & Dick, G. S. *Geology* **24,** 47–51 (1996).
3. Granger, D. E., Kirchner, J. W. & Finkel, R. C. *Geology* **25,** 107–110 (1997).
4. Finnegan, N. J., Schumer, R. & Finnegan, S. *Nature* **505,** 391–394 (2014).
5. Raymo, M. E. & Ruddiman, W. F. *Nature* **359,** 117–122 (1992).
6. Whipple, K. X. *Nature Geosci.* **2,** 97–104 (2009).
7. Sadler, P. M. *J. Geol.* **89,** 569–584 (1981).
8. Schumer, R. & Jerolmack, D. J. *J. Geophys. Res.* **114,** F00A06 (2009).
9. Crosby, B. T. & Whipple, K. X. *Geomorphology* **82,** 16–38 (2006).
10. Korup, O., Montgomery, D. R. & Hewitt, K. *Proc. Natl Acad. Sci. USA* **107,** 5317–5322 (2010).
11. Peters, S. E. *J. Geol.* **114,** 391–412 (2006).
12. Sadler, P. M. & Jerolmack, D. J. in *Strata and Time: Probing the Gaps in Our Understanding* (eds Smith, D. & Burgess, P.) (Geol. Soc. Lond. Spec. Publ., in the press).
13. Sadler, P. *Geol. Soc. Am. Abstr. Programs* **45,** 86 (2013).
14. Herman, F. *et al. Nature* **504,** 423–426 (2013).

BIRD FLIGHT

Fly with a little flap from your friends

In-air measurements of northern bald ibises flying in a V formation show that the birds conform to predictions for saving energy by regulating their relative body position and synchronizing their flapping motion. SEE LETTER P.399

FLORIAN T. MUIJRES & MICHAEL H. DICKINSON

The elegant V formations of migrating birds provide a picturesque harbinger of summer's end, but why do the birds fly in such a precise formation? There are rumours that Allied bomber pilots during the Second World War noticed that their fuel economy increased when their squadrons flew in a V formation. Although these apocryphal tales have not been confirmed, the energy-saving benefits of formation flying have been reported in both civil[1] and military[2] aviation. For example, by maintaining one wing tip in the wake of a forward plane, a fighter jet can reduce its energy consumption by up to 18% (ref. 2). However, exploiting the benefits of formation flight is more challenging for birds than for fixed-wing aircraft — birds not only need to adjust their position relative to each other, but also must synchronize their wingbeat patterns[3]. On page 399 of this issue, Portugal *et al.*[4] show that at least one bird species exhibits the requisite synchronization of body position and flapping motion to reduce energetic cost during migratory flight.

The principle by which formation flight saves energy derives from the way wings disturb the air as they move[1,5]. To create lift, wings accelerate airflow over their top surface compared with their bottom surface. Thus, relative to the still air through which they move, wings create a net circular flow of air that is directed rearward over the top surface and forward under the bottom surface. The greater the circulation a wing creates, the higher the lift it produces. At each wing tip, however, the circulation around the wing rolls up into a tip vortex, which extends backward like a tube,

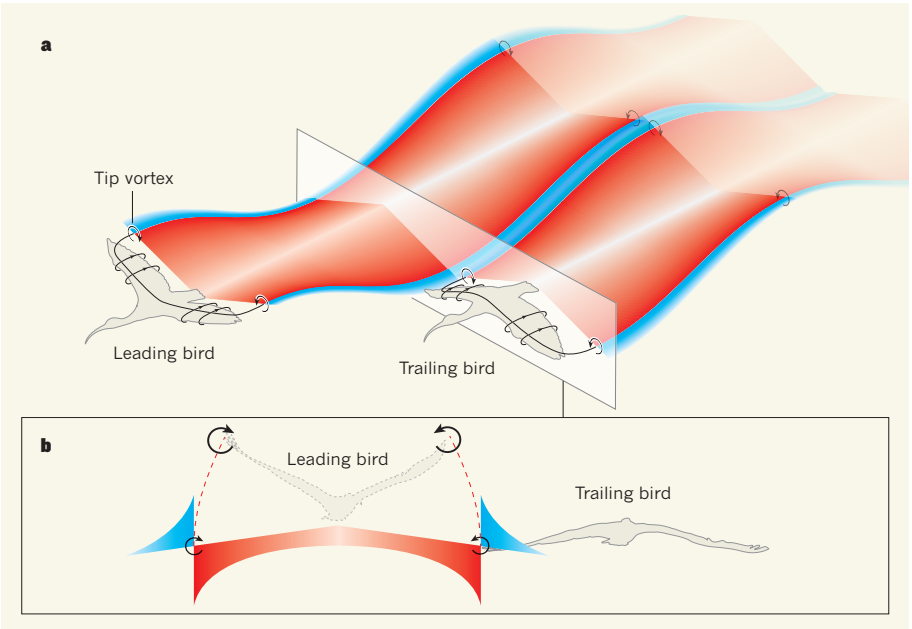


Figure 1 | Spatial synchrony. Flight creates a looping motion of air around a bird's wings; at the wing tips this circulation forms a vortex, creating air movement that extends behind the bird. Airflow down the middle of this wake is directed downwards (the downwash; red), whereas the area outside the tip vortices is a region of upwash (blue). A bird flying behind another bird experiences the aerodynamic forces of the downwash and upwash created by the leading bird (**a**, side view; **b**, rear view). Portugal and colleagues[4] show that northern bald ibises spatially synchronize their wing movements while flying in a V formation, such that the trailing bird's wing moves through the area of maximum upwash created by the leading bird. This results in more-efficient lift production and energy savings.

creating a horseshoe-shaped structure — the wake — that extends as the animal moves forward (Fig. 1a).

Owing to the laws of conservation, the circulation around the wings is equal in strength to the circulation around each of the two tip vortices. Air flows down through the middle of the wake, and the steady change in momentum as this downwash region elongates is equal to the lift created by the wings. Any animal flying behind another should avoid the downwash region, which would literally push them earthward. Just outside the downwash zone, however, there is a small region of upwash, created by the circular flow of air in the tip vortices. By careful placement of its own wing tip, a trailing bird can exploit the upwash generated by the tip vortex of a leading bird, thereby generating lift more efficiently and reducing its flight cost (Fig. 1b).

For an aeroplane pilot, keeping one wing precisely within the small upwash region of a leading plane's tip vortex is tricky enough, but for a bird the problem is further complicated by the flapping wings of its neighbours, which create tip vortices that undulate up and down. A bird that is following another bird must carefully adjust its own flapping motion, not in perfect temporal synchrony with the leader, but rather at a precise phase lag that tracks the tip vortex as it oscillates. When flying most efficiently, all the birds in a formation should flap with a precise metachrony (a wave-like synchrony), such that the flapping phase changes systematically from the leader to each bird down the line. Several theoretical studies[3,5–8] have predicted how birds flying in formation could optimize energy savings by tuning their spacing and wing motion, and geese flying in a V formation have been observed to align their body positions in a way that might save energy[7]. But until now, no experimental data have shown that birds are capable of the precise adjustment of flapping phase that is necessary to track undulating tip vortices.

Northern bald ibises (*Geronticus eremita*) often fly in a V formation when they migrate. To examine their behaviour during formation flight, Portugal and co-workers mounted custom-built data loggers on 14 ibises that accurately measured the body position and flapping dynamics of each bird. The authors found that trailing birds flew so as to keep their inner wing in the upwash zone of the bird in front of them, just as predicted by theory[3]. This requires not only correct regulation of body position, but also proper adjustment of the flapping phase, so that each bird's wing tip follows the undulating tip vortex of the individual in front of it. Because the birds occasionally shifted position within the formation, situations occurred in which trailing birds briefly flew directly behind a leading bird. In these situations, the trailing birds tended to flap their wings in strict antiphase with the leading bird, thereby minimizing the negative interactions with the downwash region of the wake. This change in behaviour indicates that the ibises actively adjust their flapping pattern under different conditions.

Although these findings are qualitatively consistent with theoretical predictions, many challenging questions remain. For example, how much energy do the birds actually save? The best existing evidence that V formations save a significant amount of energy is that pelicans have a lower heart rate and show reduced flapping frequency when flying in formation compared with when flying solo[9]. Accurate measurements of metabolic rate will be crucial for a more precise understanding of the underlying aerodynamics of formation flight and for greater insight into the ecology of bird migration. Do ibises and other birds instinctively flap in an efficient manner when flying in formation, or do they learn to adjust their body position and wing motion because it 'feels' easier? And if the strategy is so useful, why do many species of small migrating birds not fly in a V formation? Might the benefits of formation flying decrease with body size, or is the requisite control of body position and wing motion more difficult for smaller, faster-flapping birds? Although our understanding of V formations has improved, there is still much to ponder when looking skyward on late summer days. ■

Florian T. Muijres *and* **Michael H. Dickinson**
are in the Department of Biology, University of Washington, Seattle, Washington 98195-1800, USA.
e-mail: flyman@uw.edu

1. Ning, S. A. *Aircraft Drag Reduction Through Extended Formation Flight.* PhD thesis, Stanford Univ. (2011).
2. Vachon, M. J., Ray, R., Walsh, K. & Ennix, K. in *AIAA Atmos. Flight Mech. Conf. Exhib.* Abstr. 2002-4491 (Am. Inst. Aeron. Astronautics, 2002); http://dx.doi.org/10.2514/6.2002-4491
3. Willis, D. J., Peraire, J. & Breuer, K. S. *25th AIAA Appl. Aerodynam. Conf.* Abstr. 2007-4182 (2007); http://dx.doi.org/10.2514/6.2007-4182
4. Portugal, S. J. *et al. Nature* **505,** 399–402 (2014).
5. Lissaman, P. B. S. & Lundry, J. L. *J. Aircr.* **5,** 17–21 (1968).
6. Hummel, D. *J. Theor. Biol.* **104,** 321–347 (1983).
7. Badgerow, J. & Hainsworth, F. *J. Theor. Biol.* **93,** 41–52 (1981).
8. Maeng, J.-S., Park, J.-H., Jang, S.-M. & Han, S.-Y. *J. Theor. Biol.* **320,** 76–85 (2013).
9. Weimerskirch, H., Martin, J., Clerquin, Y., Alexandre, P. & Jiraskova, S. *Nature* **413,** 697–698 (2001).

ASTROPHYSICS

Black hole found orbiting a fast rotator

Stars of spectral type 'Be' are often found with neutron stars or other evolved analogues, but a black-hole companion has never been spotted before. Optical emission from a black hole's surroundings has given it away. SEE LETTER P.378

M. VIRGINIA MCSWAIN

Stellar-mass black holes, which are formed from the gravitational collapse of massive stars, are unusually scarce in our Universe. It is not yet clear whether there are fewer of them than expected or whether they are just hard to find, but either way they are deserving of their exotic reputation. Therefore, the discovery by Casares *et al.*[1], reported on page 378 of this issue, of a stellar-mass black hole orbiting around a star dubbed MWC 656 is like finding a needle in a haystack. This black hole does not emit X-ray radiation — as black holes are expected to do — so it could be the first sign of a large population of 'quiescent' black holes.

MWC 656 is itself interesting because it is surrounded by a dense outflow from the star's equator caused by a combination of its fast rotation (the projected rotational velocity[2] is about 300 kilometres per second) and pulsations that can eject material from the equator[3]. The resulting circumstellar disk produces spectral emission lines from hydrogen and other elements, meriting its classification as a 'B-emission' or 'Be' star. Fast rotation is a requirement in the formation of Be stars, and they probably acquire their high angular momentum during mass transfer from a massive companion star that eventually explodes as a supernova. In fact, many Be stars are found with highly evolved companions, usually neutron stars that are remnants of the post-supernova massive stars[4]. A few Be stars have companion stars that have been stripped down to just their helium cores[5], but such a hot object is ruled out in the case of MWC 656 because there is no significant ultraviolet-light contribution coming from anywhere other than the Be star. MWC 656 is the first Be star to have a black-hole companion detected (Fig. 1).

Casares *et al.* have identified emission lines from helium plasma that is trapped in an accretion disk around the black hole, as well as emission from the disk around the Be star, that provide a robust measurement of the mass ratio between the star and the black hole. Such emission lines are notoriously difficult to measure: they are broad and often asymmetric, complicating the usual procedures used to

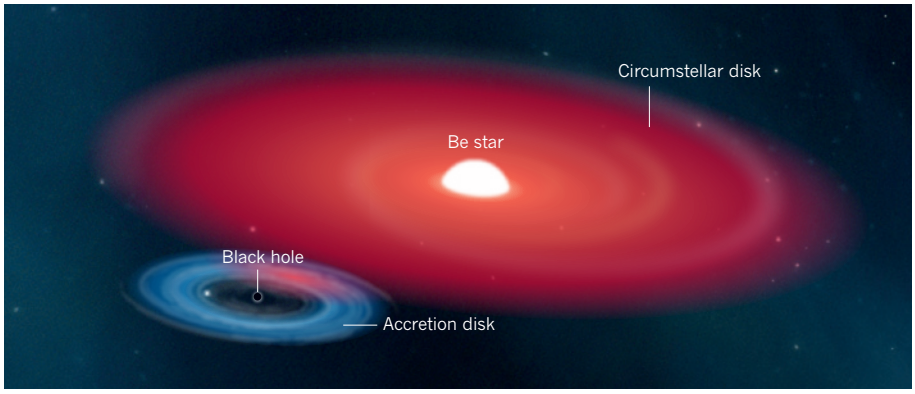


Figure 1 | A Be star with a black-hole companion. Casares *et al.*[1] have detected a quiescent black hole circling a Be star by measuring optical emission from the black hole's accretion disk and from the large disk around the star. As the black hole orbits the star, some material from the circumstellar disk is transferred to the accretion disk. However, the high angular momentum of material in the accretion disk inhibits it from falling into the black hole, so there will be a gap between the accretion disk and the black hole, and the black hole remains quiescent.

measure the line centres, but the authors have taken care to reduce any underlying systematic errors. Taken together with the mass of the Be star, which is about 10–16 solar masses, the measured ratio implies that the black hole has a mass of between 3.8 and 6.9 solar masses.

In studies of stellar evolution, conventional wisdom tells us that stellar-mass black holes form during the collapse of the cores of very massive stars — those with masses more than 25 times that of the Sun[6] — once the stars exhaust their fuel, and that the collapse is possibly accompanied by a supernova. The supernovae that massive stars (8–25 solar masses) undergo are expected to produce neutron star remnants instead. These massive stars tend to form within close groups of stars, so binary star systems are the norm, and triple and quadruple systems are not unusual. The catastrophe of a supernova in a binary has dramatic consequences: if more than half of the total-system mass is lost, or if 'kick' velocity from the explosion propels the newly formed supernova remnant with enough momentum, the remnant and companion star could fly off in opposite directions[7]. But if the companion star does remain gravitationally bound to the remnant, an X-ray binary is formed: the black hole or neutron star remnant interacts with the remaining star to produce X-ray emission.

Theorists predict that stellar-mass black holes are abundant. If this is so, we should find them all over the Milky Way. Many of them ought to be bound in X-ray binaries, whereas others should be freely floating through space. There are probably tens of millions of massive stars in the Milky Way that could potentially collapse into black holes, but there are only about 50 stellar-mass black holes known with good confidence[8]. X-ray studies of young star-forming regions such as the Carina Nebula, which might contain at least a few recent supernovae products, have not found any black holes[9]. Even large sky surveys are coming up with little as they search for the subtle brightness variations of stars whose light bends around a foreground black hole that passes in front of the star (an effect known as microlensing)[10].

MWC 656 presents a rare opportunity to study mass transfer, angular momentum and accretion-disk physics around a quiescent black hole. Casares *et al.* find a hint of a hotspot on the black hole's accretion disk that suggests that mass is pulled away from the Be star's disk, crashing into the accretion disk when the stars make their closest approach during their orbit around one another. The absence of X-ray emission from this system is evidence that material is not channelled into the black hole; rather, it must be retained in a holding pattern within the accretion disk. Gas in the outer regions of the Be star's disk will have high angular momentum, which will be transferred to the accretion disk during the mass transfer. Without an efficient mechanism to remove this angular momentum, accretion will be suppressed and the black hole will remain quiet. If there exists a larger population of Be star–black-hole binaries, such quiescence is probably the rule, not the exception. Casares *et al.* have shown us a way to find them. ■

M. Virginia McSwain *is in the Department of Physics, Lehigh University, Bethlehem, Pennsylvania 18015, USA.*
e-mail: mcswain@lehigh.edu

1. Casares, J. *et al. Nature* **505,** 378–381 (2014).
2. Williams, S. J. *et al. Astrophys. J.* **723,** L93–L97 (2010).
3. McSwain, M. V., Huang, W., Gies, D. R., Grundstrom, E. D. & Townsend, R. H. D. *Astrophys. J.* **672,** 590–603 (2008).
4. Liu, Q. Z., van Paradijs, J. & van den Heuvel, E. P. J. *Astron. Astrophys.* **455,** 1165–1168 (2006).
5. Gies, D. R. *et al. Astrophys. J.* **493,** 440–450 (1998).
6. Fryer, C. L. *Astrophys. J.* **522,** 413–418 (1999).
7. Brandt, N. & Podsiadlowski, P. *Mon. Not. R. Astron. Soc.* **274,** 461–484 (1995).
8. Belczynski, K., Wiktorowicz, G., Fryer, C. L., Holz, D. E. & Kalogera, V. *Astrophys. J.* **757,** 91 (2012).
9. Hamaguchi, K. *et al. Astrophys. J.* **695,** L4–L9 (2009).
10. Moniez, M. *Gen. Relativ. Gravit.* **42,** 2047–2074 (2010).

50 Years Ago

Space Carrier Vehicles by Oswald H. Lange and Richard J. Stein — The book begins with a useful conspectus of United States space launching missiles from the *Juno* 1 to the *Saturn C*-5: a list of their achievements is included, which shows the *Thor-Agena B* to be well in the lead, with 39 successful launches before the end of 1962 … Further subjects discussed are inertial guidance and control, the fabrication of the missiles (including an informative series of photographs of the Saturn vehicle under construction) … and, finally, the layout and construction of launching sites (with photographs of the Saturn launch complex at Cape Kennedy) … The book shows a bias in favour of German or American achievements: p.1 gives the impression that the first satellite was launched by the United States.
From *Nature* 18 January 1964

100 Years Ago

An article in *Engineering* for January 9 directs attention to the waning supply of petroleum. Although a continually greater supply of petroleum is being placed on the market, this increased output is secured only by sinking more wells and boring to a greater depth, showing that the surface supply is becoming exhausted. At the beginning of this century the wells touched 1100 ft., and to-day the average level of the oil may be placed at 2000 ft. — an ominously rapid rate of sinking … America, by reckless expenditure of her resources, has increased her annual output to 200 million barrels, yet the demand for oil for special purposes has become so great that the rise in price is considerable — so great, indeed, that competition with coal for ordinary purposes has become impossible.
From *Nature* 15 January 1914

PLANT SCIENCE

Fairy chemicals

Those who cultivate manicured lawns curse 'fairy rings' of mushrooms (pictured) and the rapid grass growth associated with them. The compound that stimulates this growth, 2-azahypoxanthine (AHX), was isolated from a fungus in 2010, but Choi *et al.* now report in *Angewandte Chemie* that plants also produce it (J.-H. Choi *et al. Angew. Chem. Int. Edn* http://doi.org/f2phhs; 2014).

The authors treated several plants with AHX, and observed that it was metabolized to a compound called 2-aza-8-oxohypoxanthine (AOH). They went on to show that both AHX and AOH are produced by plants, and are present in rice at levels similar to those of plant hormones.

Choi *et al.* found that a member of the purine metabolic pathway is converted to AHX and AOH in rice, and they extracted crude enzymes that catalyse the reactions involved from rice and *Arabidopsis*, a model plant. They conclude that AHX and AOH are formed in a previously unknown metabolic pathway.

Intriguingly, AOH stimulates rice growth, albeit not as much as AHX. Frustrated haters of fairy rings could perhaps take heart from the thought that both compounds hold promise for horticulture. **Andrew Mitchinson**

WALLY EBERHART/VISUALS UNLIMITED/CORBIS

MOLECULAR BIOLOGY

The tug of DNA repair

The transcription enzyme RNA polymerase stalls at DNA lesions, hindering their repair. Accessory factors dislodge the enzyme by pushing it forwards, but a study finds that pulling it backwards may also be effective. **SEE ARTICLE P.372**

IRINA ARTSIMOVITCH

DNA damage caused by genotoxic agents, from solar ultraviolet light to free radicals generated during cellular metabolism, is unavoidable. Unless repaired, such damage may directly threaten cell survival or lead to mutations and disease. All organisms therefore rely on repair systems to maintain DNA integrity, with many components of those systems highly conserved throughout evolution. In this issue, Epshtein *et al.*[1] (page 372) describe how two enzymes — an RNA polymerase and UvrD, a helicase — cooperate to target a damaged site for repair.

RNA polymerase (RNAP) carries out the essential job of transcribing DNA into RNA. When RNAP runs into a lesion in the transcribed (template) strand, it backtracks along the DNA, but still occludes the damaged site (Fig. 1a, b), an effect that might be expected to hinder DNA repair. Unexpected observations[2] that, *in vivo*, damage is preferentially repaired in transcribed rather than non-transcribed regions led to the discovery of transcription-coupled repair (TCR), a branch of the nucleotide-excision repair (NER) pathway that removes many lesions, including ultraviolet-induced thymine dimers[3].

NER of naked DNA in bacteria involves a complex of Uvr proteins: UvrA and UvrB recognize the lesion; UvrC excises the damaged segment; and UvrD helicase displaces it. The resulting gap is then enzymatically repaired by two enzymes, DNA polymerase and a ligase. In TCR, RNAP takes over the damage-recognition task, but has to be removed to allow the Uvr complex to access the lesion.

In a broadly accepted model for TCR, a protein known as Mfd pushes the stalled RNAP forwards (Fig. 1c), releasing it from the DNA[4], and recruits UvrA[2], thereby directly coupling transcription to repair. Intriguingly, bacterial strains lacking Mfd are quite resistant to damage caused by ultraviolet light[2], implying that alternative means of repair exist. Epshtein *et al.* describe a TCR pathway of opposite polarity to the conventional mechanism, wherein UvrD pulls RNAP even further backwards, exposing the lesion for repair (Fig. 1d).

The role of UvrD in NER is well established — it acts late in the pathway, after RNAP has dissociated from the DNA and the UvrABC proteins have processed the lesion. However, Epshtein and colleagues observed that, in the bacterium *Escherichia coli*, UvrD interacts with RNAP as frequently as the general transcription factors NusA and NusG. This finding prompted the authors to re-examine the role of UvrD in TCR.

Their analysis convincingly demonstrates that UvrD forms a binary complex with RNAP and induces it to backtrack both *in vivo* and in an *in vitro* system. The researchers also found that UvrD relieves an RNAP-imposed block to UvrABC excision of a thymine dimer in a minimal NER system reconstituted *in vitro*.

Cells lacking UvrD are highly sensitive to several genotoxic agents that can induce NER, but Epshtein *et al.* observed that this effect is reversed when the two known anti-backtracking processes are inhibited so that RNAP can retreat even in the absence of UvrD. Moreover, the authors found that, *in vitro*, UvrD activities require an energy source (ATP nucleotides), consistent with the enzyme's function as a motor protein.

UvrD translocates on a single DNA strand to displace DNA-bound proteins, an activity thought to be mechanistically distinct from its helicase activity, which separates (melts) the two DNA strands[5]. The melted non-template DNA strand is a target for several transcriptional regulators. Could UvrD slide on the non-template DNA to push on RNAP?

Consistent with this model, Epshtein *et al.* showed that UvrD interacts with the non-template strand. They then sought to identify amino-acid residues of UvrD and RNAP that make contacts between the two molecules, using a powerful experimental approach[6] that allows the construction of a high-confidence three-dimensional map of the complex, even though a structure of the complex has not been determined. This analysis identified the β-subunit flap domain of RNAP as a contact site for UvrD. The mapping results place UvrD at the upstream end of the transcription bubble, the structure that forms when part of the DNA helix is unwound by RNAP (Fig. 1). This is a logical location for pushing the RNAP backwards, and may explain the involvement of NusA — which also interacts with the β flap — in NER[7]. Indeed, Epshtein and colleagues show that NusA cooperates with UvrD *in vitro* and may favour backtracking *in vivo*.

UvrD was first implicated in DNA repair more than 40 years ago (see ref. 8, for example), but even the most basic questions about its role remain matters of debate. It is not known whether UvrD functions as a helicase or a translocase to carry out its diverse functions, or as a monomer or a dimer. The current study does not answer any of these questions, and instead poses a few more; for example, does UvrD bind only to an RNAP stalled at a lesion, and how does it compete with other regulators that bind to the non-template DNA? It also calls for textbooks to be revised: rather than (or in addition to) displacing the damaged DNA segment, UvrD displaces the lesion-stalled RNAP. This conceptual shift will undoubtedly reignite interest in UvrD, which should, in turn, help to address the outstanding questions.

The UvrD-dependent TCR pathway seems to be dominant in *E. coli*, but is it universally conserved? It relies on backtracking, a mechanism widely used by cellular RNAPs to regulate gene expression[9]. Yeast pol II, the RNAP that transcribes most yeast genes, is structurally similar to the bacterial enzyme, retreats at lesions and associates with Ssl2, a homologue of human XPB helicase, which has the same translocation polarity as UvrD and is essential for NER. Mutations in XPB are linked to xeroderma pigmentosum, a human disorder associated with a high risk of developing skin cancer.

But, unlike UvrD, XPB is thought to melt and load the DNA into RNAP. It also binds at the 'wrong' place, contacting the downstream DNA and a transcription factor (TFIIE) that occupies the upstream part of the would-be bubble in the transcription complex when it is poised to start synthesis[10]. TFIIE subsequently dissociates from RNAP but is replaced by a NusG-like factor, which forms a processivity clamp[11] (a structure that enables uninterrupted RNA synthesis in all domains of life). NusG homologues bind to the non-template strand and would be expected to exclude XPB (or UvrD). Could the clamp be lost at a lesion? Dissecting the interplay between these and other regulators will require the powerful mapping approaches that helped Epshtein and colleagues to establish UvrD as a key factor for TCR. ■

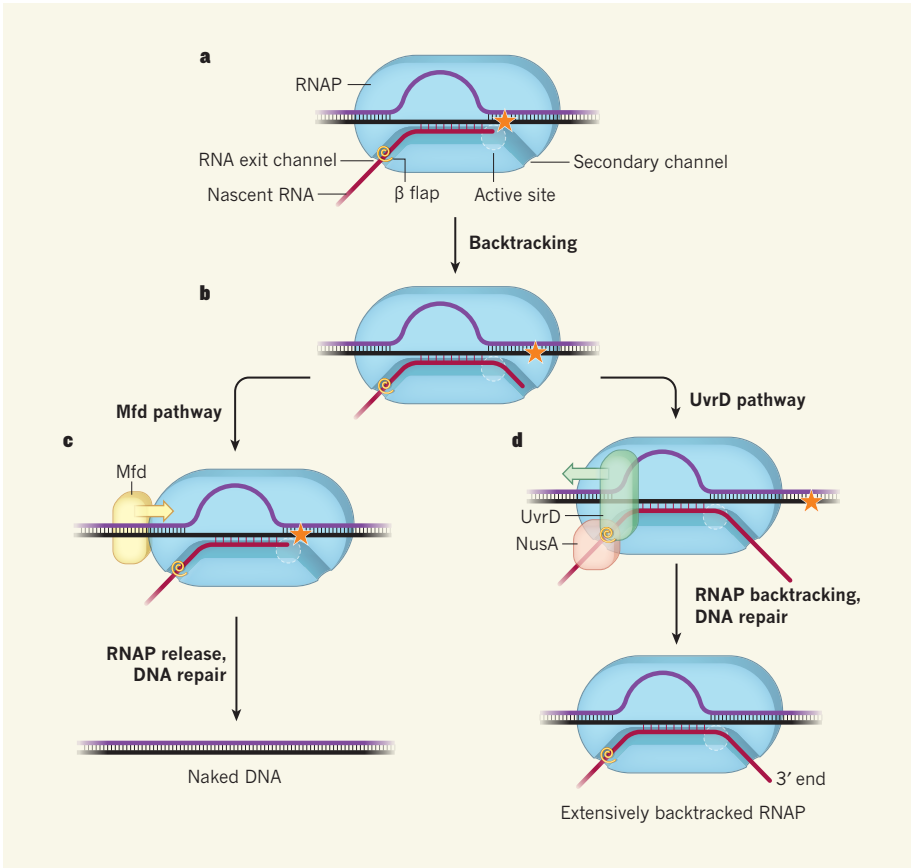


Figure 1 | Alternative routes for transcription-coupled repair in bacteria. **a**, **b**, In a transcription complex, the two DNA strands separate to form a bubble; the non-template strand (purple) is exposed on the surface of RNA polymerase (RNAP); the β flap forms part of the RNA exit channel. When RNAP encounters a DNA lesion (orange star), it backtracks, threading the nascent RNA into the secondary channel and blocking access to the lesion. **c**, Forward translocation of RNAP induced by sliding of the Mfd protein (yellow arrow) on the double-stranded DNA releases RNAP, allowing repair and generating naked, intact DNA. **d**, Epshtein *et al.*[1] report that backward translocation (green arrow) induced by UvrD, which binds to the non-template strand and the β flap, exposes the lesion; the transcription factor NusA may stabilize the UvrD–β-flap interaction. Following repair, the arrested RNAP must be reactivated to position the 3′ RNA end in the active site, or removed (for example, by Mfd).

Irina Artsimovitch *is in the Department of Microbiology and the Center for RNA Biology, Ohio State University, Columbus, Ohio 43210, USA.*
e-mail: artsimovitch.1@osu.edu

1. Epshtein, V. *et al. Nature* **505,** 372–377 (2014).
2. Selby, C. P. & Sancar, A. *Science* **260,** 53–58 (1993).
3. Ganesan, A., Spivak, G. & Hanawalt, P. C. *Prog. Mol. Biol. Transl. Sci.* **110,** 25–40 (2012).
4. Roberts, J. & Park, J.-S. *Curr. Opin. Microbiol.* **7,** 120–125 (2004).
5. Yang, W. *Annu. Rev. Biophys.* **39,** 367–385 (2010).
6. Rappsilber, J. *J. Struct. Biol.* **173,** 530–540 (2011).
7. Cohen, S. E. *et al. Proc. Natl Acad. Sci. USA* **107,** 15517–15522 (2010).
8. Ogawa, H., Shimada, K. & Tomizawa, J.-I. *Mol. Gen. Genet.* **101,** 227–244 (1968).
9. Nudler, E. *Cell* **149,** 1438–1445 (2012).
10. Grünberg, S., Warfield, L. & Hahn, S. *Nature Struct. Mol. Biol.* **19,** 788–796 (2012).
11. Werner, F. *J. Mol. Biol.* **417,** 13–27 (2012).

This article was published online on 8 January 2014.

nature INSIGHT

FRONTIERS IN BIOLOGY

16 January 2014 / Vol 505 / Issue No 7483

Cover illustration
Nik Spencer

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

nature publishing group

The Insight 'Frontiers in biology' aims to cover timely and important developments in biology, ranging from the subcellular to the organismal level, and including molecular mechanisms and biomedicine.

The collection begins with a discussion of state-of-the-art cancer-predisposition-gene discovery. Nazneen Rahman shows how these genes can provide insight into the mechanisms of cancer causation, and how their clinical use has had a substantial impact on diagnosis, optimized management and cancer prevention.

Next, Karl Deisseroth discusses how analysing brain function at the level of neural projections can be a fruitful approach towards understanding the basis of behaviours relevant to psychiatric diseases. Technical advances have allowed perturbation of the activity of specific components of neural circuits, enabling causal testing of their functions.

Adam Kepecs and Gordon Fishell go on to consider the organization of interneurons, which perform a crucial role in regulating inhibition within neural circuits. Owing to their diverse morphology, connectivity and physiology, interneurons have by and large defied generalized classification. Here, the authors explore data supporting an organization in which the cells are defined by a developmentally specified set of cardinal classes.

Sean Morrison and David Scadden highlight recent progress in our understanding of the haematopoietic stem-cell niche in bone marrow, and discuss how studying the involvement of the microenvironment in normal and disease physiology promises new approaches to treat haematological disorders.

Further advances, this time in genetics and systems-based approaches, have resulted in a renaissance in mitochondrial research. Jonathan Friedman and Jodi Nunnari discuss how this resurgence is both redefining and extending our knowledge about mitochondrial behaviour and communication.

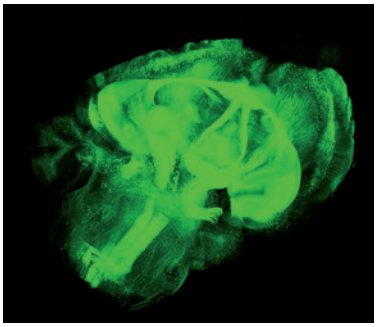
In the penultimate Review, Pier Paolo Pandolfi and colleagues discuss the ways in which various forms of endogenous RNA species can interact with and alter the expression or stability of other RNAs.

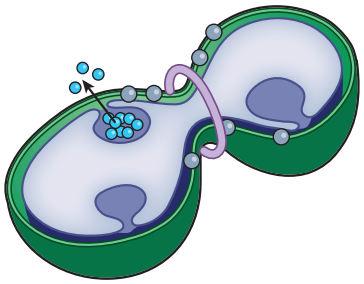
Finally, Robert Ross, Molly Weiner and Haifan Lin discuss recent studies that hint at the intriguing roles of piRNAs — originally considered to be germline-specific regulatory RNAs — in somatic cells.

Francesca Cesari, I-han Chou, Angela K. Eggleston, Noah Gray, Deepa Nath, Sadaf Shadan & Magdalena Skipper
Senior Editors

CONTENTS

REVIEWS





REVIEW

doi:10.1038/nature12981

Realizing the promise of cancer predisposition genes

Nazneen Rahman[1,2]

Genes in which germline mutations confer highly or moderately increased risks of cancer are called cancer predisposition genes. More than 100 of these genes have been identified, providing important scientific insights in many areas, particularly the mechanisms of cancer causation. Moreover, clinical utilization of cancer predisposition genes has had a substantial impact on diagnosis, optimized management and prevention of cancer. The recent transformative advances in DNA sequencing hold the promise of many more cancer predisposition gene discoveries, and greater and broader clinical applications. However, there is also considerable potential for incorrect inferences and inappropriate clinical applications. Realizing the promise of cancer predisposition genes for science and medicine will thus require careful navigation.

Genetic predisposition to cancer has been recognized for centuries, initially through observation of unusual familial clusterings of cancer. In 1866, neuroanatomist Paul Broca published one of the earliest reports on the subject, detailing a striking history of breast cancer in 15 members of his wife's family[1]. Broca, controversially for the time, proposed that this was evidence of hereditary predisposition to cancer. Fifty years later, biologist Theodor Boveri published his visionary theory that somatic acquisition of "particular, incorrect chromosome combinations" underlies cancer. His paper was equally prophetic about inherited predisposition to cancer, predicting it could result from "weakened resistance against the action of factors that stimulate cell division"[2]. In 1971, mathematical modelling of the epidemiology of retinoblastoma by geneticist Alfred Knudson suggested a 'two-hit' model whereby both alleles of a specific gene were required to be inactivated for retinoblastoma to occur[3], thus echoing Boveri's predictions. In 1987, the retinoblastoma predisposition gene *RB1* was discovered, and in hereditary cases one allele was mutated in the germ line with the second allele inactivated somatically[4].

There is no definitive definition of a cancer predisposition gene (CPG). For the purposes of this Review, I have restricted inclusion to those genes in which rare mutations confer high or moderate risks of cancer (greater than twofold relative risks) and to those for which at least 5% of individuals with the relevant mutations develop cancer. For most genes, both the risks and penetrance are considerably higher than these minimum criteria. Common variants conferring very small increases in risk discovered through genome-wide association studies (GWAS) are not included within this definition. Such variants are important components of the genetic architecture of cancer and are addressed in other reviews[5–7]. Through extensive literature and database evaluations, I identified 114 CPGs that form the basis of this Review (Fig. 1; see Supplementary Information).

Conventionally, cancer predisposition reviews have focused on select cancers and/or sets of genes. In an era of whole-genome sequencing, information about all known CPGs is increasingly desirable in both research and clinical practice. Here, I have aspired to integrate knowledge from three decades of research to provide a distillation of key CPG characteristics. I also discuss the exciting prospects and potential pitfalls of future discoveries and clinical applications.

Discovery of CPGs

The 114 CPGs were discovered over the past 30 years through multiple strategies (Fig. 2). Since 1990, at least one new CPG has been identified each year, peaking in 1996, when 10 CPGs were reported[8–16]. Genome-wide linkage analysis, an agnostic approach that allows tracking of disease-associated genomic markers in high penetrance familial clusters has been the most successful strategy, yielding 59 CPGs, mostly during the 1990s when the methodology became routine. Next-generation sequencing is leading to a new crop of CPGs being discovered through genome-wide mutational analyses such as exome and genome sequencing[17,18].

The remaining genes were identified through various candidate-based strategies. Large numbers of candidate CPGs have been proposed and investigated. For most, no association with cancer predisposition has been found. This perceived failure led to candidate-based approaches falling out of favour. However, certain strategies have proved very successful both as stand-alone discovery methods and in facilitating linkage studies. Candidates pursued as surrogates of cancer predisposition — for example distinctive cellular phenotypes such as defective DNA repair, mosaic aneuploidies or telomere shortening — have contributed to the discovery of many CPGs[9,10,19,20]. Genetic-pathway candidates, that is, genes selected because they function in similar pathways to known CPGs, have also yielded new predisposition genes, particularly in colorectal, breast, ovarian and endocrine cancers[21–28]. Candidate genes that have been chosen because they are somatically mutated in cancers have, perhaps surprisingly, led to the identification of only 12 CPGs[29–40].

Overlap of somatic and germline cancer genes

It is interesting and instructive to consider the overlap between the known germline and tumour-mutated cancer genes. Currently, the COSMIC (Catalogue of Somatic Mutations in Cancer) database includes 468 genes that are somatically mutated in cancers[41]. Of these, 49 are also known to be CPGs. Conversely, 65 of the 114 CPGs are known to be somatically mutated. These data imply that 10% of somatically mutated cancer genes also confer susceptibility to cancer when mutated in the germ line, but that 40% of germline-mutated CPGs can also contribute to oncogenesis when mutations occur only in tumours (Fig. 3). This apparent discrepancy is, at least in part, an artefact of different research approaches; it is common for the frequency of somatic CPG mutations to be investigated, but it is unusual for somatically mutated genes to be evaluated for their role in cancer predisposition. The latter has been exacerbated by cancer genome

[1]Division of Genetics and Epidemiology, Institute of Cancer Research, London SW2 5NG, UK. [2]Cancer Genetics Unit, Royal Marsden Hospital Foundation Trust, London SM2 5PT, UK.

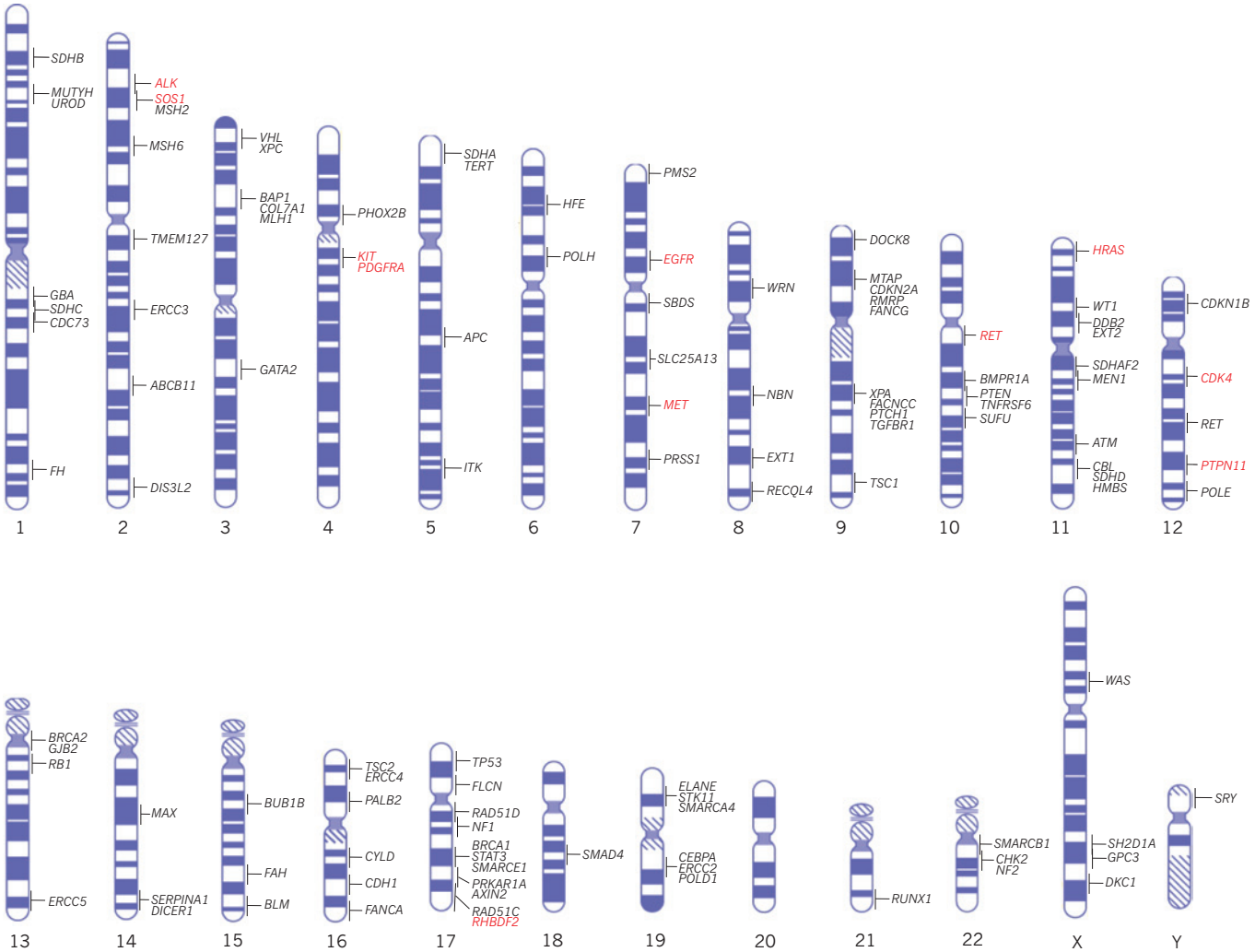


Figure 1 | Chromosomal locations of 114 cancer predisposition genes. Gain-of-function mutations in genes that predispose carriers to cancer are shown in red. Loss-of-function mutations in genes that predispose carriers to cancer are shown in black.

sequencing studies in which a key filtering step is the removal of variants that are present in normal tissue to focus attention on the potential cancer-driving mutations that are present only in the tumour[42]. Thus, it is highly likely that there is an underestimate of the overlap between somatic cancer driver mutations and germline cancer predisposing mutations, and mutual interrogation of such genes could prove to be a useful approach for identification of new cancer genes.

Overlap of high and low penetrance cancer-associated variants

One notable absentee in the successful CPG discovery strategies is identification through GWAS candidature. There are 391 known common variants that confer small increased risks of cancer[43]. The underlying causal gene or mechanism has only been identified for a small minority, but none have been shown to be sentinels of rare and higher penetrance mutations in new CPGs. Furthermore, only 15 SNPs (single nucleotide polymorphisms) are within known CPGs and none are associated with cancers that occur in carriers of rare, penetrant mutations. For example, *ATM* D1853N (the only exonic cancer GWAS variant in a CPG) confers a protective effect against melanoma, but it is not associated with the cancers that occur in biallelic or monoallelic *ATM* mutation carriers[44,45]. Similarly, the SNP rs78378222 alters the polyadenylation signal of *TP53* and is associated with various cancers, but not those that typically occur in patients with the cancer predisposition condition Li-Fraumeni syndrome, which is due to germline *TP53* mutations[46]. Multiple variants in the vicinity of *TERT* have been associated with several different cancers, but again not those that occur in the multisystem disorder dyskeratosis congenita, which is a recessive condition due to germline exonic *TERT* mutations[47–49]. These data suggest the mechanisms underlying the association of rare, high penetrance alleles and common, low penetrance alleles with cancer are largely distinct. This differs from several other common, complex conditions that show considerable overlap between these components of the genetic architecture[50,51].

Characteristics of CPGs

The 114 genes are located throughout the genome with little evidence of chromosomal clustering (Fig. 1).

Inheritance and mechanisms of oncogenesis

The inheritance pattern of cancer predisposition is varied; it is autosomal dominant for 65 CPGs, autosomal recessive for 28, X-linked for 4 and Y-linked for 1. Sixteen genes cause phenotypes in both monoallelic and biallelic mutation carriers, that is, they cause autosomal dominant and autosomal recessive conditions. For half of these, the recessive condition is a more severe manifestation of the dominant condition. For example, biallelic *BRCA2*, *PALB2*, *MLH1*, *MSH2*, *MSH6* and *PMS2* mutation carriers have a high risk of childhood cancer, whereas monoallelic mutation carriers have an increased risk of adult cancers[52]. For other genes in which monoallelic and biallelic mutations cause clinical phenotypes, such as *FH* and *SDHA*, cancer has not been reported in biallelic mutation carriers. This may be because of early mortality from other causes.

Most CPGs act as tumour suppressor genes with mutations that abrogate their function, promoting oncogenesis. Only 11 genes predispose

to cancer through gain-of-function mutations. Several of these, such as *RET*, *MET*, *KIT* and *ALK*, encode kinases that are rendered constitutively active by cancer predisposing mutations[53]. Much more diversity exists in the types, functions and mechanisms of oncogenesis of CPGs that when inactivated increase the risk of cancer. Many are classic tumour suppressor genes, requiring both alleles to be inactivated, but haploinsufficiency and dominant-negative mechanisms also occur. For some genes, different mutations operate through different mechanisms and lead to distinct phenotypes[54,55]. For many genes, the clinical phenotypes and cancer risks associated with CPG mutations are also influenced by other factors, both genetic and non-genetic.

Functions of CPGs

CPG research has directly resulted in insight into basic biological pathways and gene function. Several essential genes were first isolated because germline mutations within them predispose cancer, as is transparent in their names (for example, breast cancer 1, *BRCA1*; or retinoblastoma 1, *RB1*), which reflect the clinical phenotypes that led to their identification. The functions of these genes were only subsequently elucidated, often directly because of research into their role in cancer predisposition.

CPGs have a very broad range of functions. Many are ubiquitously expressed and participate in fundamental processes such as DNA repair and cell-cycle regulation. One of the enduring conundrums of cancer predisposition is why, and how, perturbation of universal cellular functions can cause exquisitely specific cancer phenotypes. However, some genes have organ-specific functions that are transparently related to the cancers with which they are associated. For example, mutations in *SLC25A13*, *ABCB11*, *FAH*, *HMBS* and *UROD* all lead to hepatic overload, liver cirrhosis and, hence, an increased risk of hepatocellular carcinoma[56].

Meaningful evaluation of the 114 CPGs for functional associations is currently precluded because so many were identified owing to their functional relationships with known CPGs. Nonetheless, some noteworthy functional networks are emerging as important in cancer predisposition, in addition to well-recognized pathways such as DNA repair. Among these are the SWI–SNF chromatin-remodelling pathway, which has been linked to rhabdoid tumours and meningiomas; the succinate dehydrogenase enzyme complex, which is associated with phaeochromocytoma and paragangliomas; and the PI(3)K–mTOR signalling pathway, which has links with several CPGs, including *TSC1*, *TSC2*, *PTEN*, *STK11*, *FLCN*, *HRAS* and *TMEM127*, and hence is associated with diverse cancers[57–59].

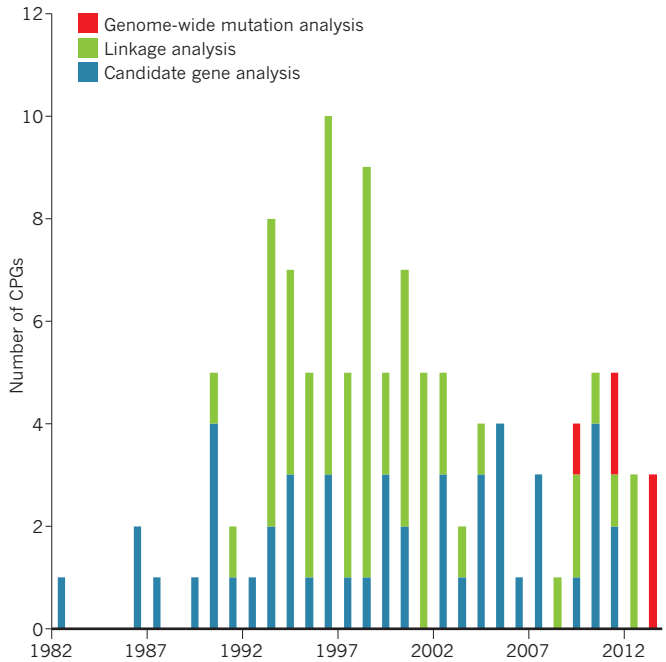


Figure 2 | Timeline of cancer predisposition gene (CPG) discovery. CPGs were first identified by candidate gene approaches (blue). As genome-wide linkage analysis (green) became increasingly routine, this led to the discovery of many CPGs. Genome-wide mutation analyses, such as exome and genome sequencing (red), are leading to the discovery of a new crop of CPGs. Since 1990, at least 1 CPG has been identified each year, peaking in 1996 when 10 CPGs were identified.

Cancer phenotypes

It is currently estimated that around 3% of cancers are due to CPG mutations, which is equivalent to more than 300,000 cancers per year worldwide. This is an underestimate because the contribution of known genes has been poorly characterized and not all genes have been identified. The contribution to individual cancers is highly variable. The highest attribution is to childhood embryonal tumours such as retinoblastoma and pleuropulmonary blastoma which are often due to germline mutations in *RB1* and *DICER1*, respectively[29,60]. This simplicity is not applicable to all childhood cancers; the embryonal kidney cancer Wilms tumour is associated with several CPGs and other predisposition mechanisms, which together account for less than 5% of cases[61,62]. At the other end of the spectrum, known CPGs make a very small contribution to some adult cancers, such as prostate and lung cancer. However, germline CPG mutations in multiple genes predispose carriers to other adult cancers such as breast, colorectal, melanoma and ovarian cancer. For some, the overall contribution of CPGs is sizeable, with around 15% of ovarian cancers, about 20% of medullary thyroid cancers and more than 30% of phaeochromocytomas due to CPG mutations[63–65].

Some CPGs preferentially predispose carriers to specific histological subtypes of a cancer. For example, *BRCA1* is particularly associated with triple-negative breast cancer and serous ovarian cancer, whereas *CDH1* is particularly associated with lobular breast cancer and diffuse gastric cancers[66,67]. The genomic profiles of cancers arising in individuals with germline CPG mutations can also be distinctive; chromothripsis, which describes localized chromosomal shattering, is striking in medulloblastomas that occur in *TP53* germline mutation carriers, and cancers in *BRCA1* or *BRCA2* mutation carriers have a characteristic mutational signature that includes substantial numbers of deletions with overlapping microhomology at the breakpoint junctions[68,69].

Non-cancer phenotypes

Clinical phenotypes in addition to cancer often occur in individuals with CPG mutations, with 87 CPGs being associated with non-cancer clinical features. These are often more discriminating and more common than cancer, and can be crucial to clinical diagnosis of the underlying cancer syndrome. The spectrum of additional clinical features is very broad. Skin manifestations are the most frequent and can be specific to the relevant CPG. They include hypopigmented and/or hyperpigmented areas, freckling, rashes, blistering, hypertrophy, skin tags, and nodules and/or lumps. Neurological, dysmorphic and skeletal manifestations also occur, but are usually nonspecific features such as microcephaly, macrocephaly, short stature and/or developmental delay. The proportion of CPGs associated with non-cancer clinical phenotypes is probably an overestimation of the true proportion, as identification of genes that result in a readily clinically recognizable phenotype is inevitably more tractable.

Genotype–phenotype associations

One of the illuminating outcomes of CPG research has been increased knowledge about the diversity of mutational mechanisms and their relationship with phenotype. Even superficially straightforward associations can mask profound complexity. For example, gain-of-function germline *HRAS* missense mutations cause enhanced MAPKK and PI(3)K signalling similar to somatic mutations[70]. However, the spectrum of germline and somatic mutations differs, and germline *HRAS* mutations not only predispose to cancer but also cause a multisystem disorder called Costello syndrome[71]. This condition includes distinctive facial dysmorphism and a wide range of cardiac, dermatological, musculoskeletal and developmental abnormalities. The role of HRAS in these processes, and why *HRAS*

mutations lead to such a complex phenotype, is unknown.

The genotype–phenotype relationships of some CPGs are extraordinarily intricate and hint at deep uncharted complexities in gene function. Most *WT1* mutations predispose carriers to Wilms tumour, genitourinary abnormalities and renal dysfunction, the severity of which is influenced by the mutational type. However, intronic mutations that alter the relative abundance of WT1 isoforms cause a distinct condition, called Frasier syndrome, which includes gonadoblastoma rather than Wilms tumour, a focal-segmental nephropathy and severe gonadal dysgenesis, which can manifest as complete sex reversal[61].

TERT, which encodes telomerase, is another example. Recently, activating promoter mutations that result in increased telomerase expression were shown to predispose carriers to melanoma[72]. Monoallelic and biallelic, primarily missense *TERT* mutations cause dyskeratosis congenita, which is characterized by various physical abnormalities, pulmonary fibrosis, bone-marrow failure and increased incidence of acute myelogenous leukaemia and squamous carcinomas of the head and neck and anogenital region[48]. Furthermore, various common SNPs in the vicinity of *TERT* confer small risks of several cancers, including breast, colorectal, testicular and prostate[47,49]. The mechanisms underlying the diversity of *TERT* phenotypic associations are unknown.

TGFBR1, which encodes a transmembrane serine/threonine kinase receptor, is one of the most extreme examples of genotype–phenotype diversity. Missense mutations in the kinase domain cause marfanoid vasculopathies with no increased risk of cancer. Truncating mutations in the same kinase domain, or missense mutations in the extracellular ligand-binding domain, cause a highly unusual condition called multiple self-healing squamous epithelioma. Individuals with the condition develop squamous-carcinoma-like locally invasive skin tumours that grow rapidly for a few weeks then spontaneously regress and scar[73].

CPG cancer risks

There is a deep and widely underappreciated complexity in the risks of cancer conferred by CPG mutations. A specific CPG mutation can confer different risks of developing different cancers. Different CPG mutations can confer different risks of developing a particular cancer. A specific CPG mutation can even confer different risks of developing a particular cancer in different contexts. *BRCA2* illustrates all of these scenarios. Loss-of-function *BRCA2* mutations confer substantial increased lifetime risks of breast and ovarian cancer but only small increased lifetime risks of prostate and pancreatic cancer[74]. However, not all mutations confer the same risk, despite most being protein-truncating mutations predicted to result in nonsense-mediated RNA decay and thus to be functionally equivalent. Loss-of-function *BRCA2* mutations in the central part of the gene confer significantly higher relative risks of ovarian cancer compared with breast cancer than mutations at either end[75]. The mechanistic basis for this highly unusual pattern is unknown. The degree of family history also impacts on the risk of cancer of *BRCA2* mutations. The lifetime breast cancer risk of female *BRCA2* mutation carriers with a strong family history is around 80%, but the risk is only about 45% for relatives of breast cancer cases unselected for family history[76,77]. This reflects, at least in part, additional modifying factors within familial clusters that increase cancer risk. Some genetic and non-genetic modifying factors of *BRCA2* and *BRCA1* cancer risks have already been identified, although it is likely there is still much to be discovered[78–81].

TP53 is another gene that has been known for more than 20 years to predispose carriers to cancer, but our knowledge of the associated cancer risks is still lamentably incomplete. All germline *TP53* mutations are typically assumed to be highly penetrant, but the widely quoted cancer lifetime risks are derived from small series of highly selected cases[82]. In fact, there is strong evidence of high variability in the types and risks of cancer associated with different *TP53* mutations. This is exemplified by *TP53* R337H, which confers a modest 10% risk of adrenocortical cancer and is not associated with increased risks of other classic Li-Fraumeni syndrome cancers such as breast cancer or sarcoma[83].

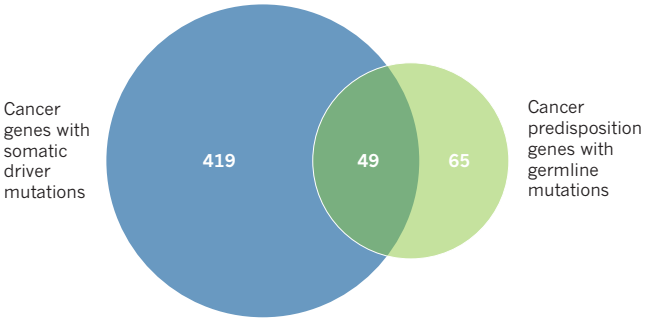


Figure 3 | Overlap between somatically mutated cancer genes and cancer predisposition genes (CPGs). 468 genes with somatic driver mutations in cancers are recorded in the COSMIC database of which 49 are also included within the 114 CPGs.

Clinical utility of CPGs

The identification of CPGs has had substantial clinical impact. Cancer is one of the foremost diseases for which such discoveries have transformed medical care in multiple areas, including cancer prevention.

Diagnosis and patient management

The benefits of determining whether a cancer is due to a germline CPG mutation are incontrovertible. As such, CPG testing has become standard for many genes, albeit typically only in highly selected cases. From a patient perspective, simply having a better understanding of why their cancer occurred is usually highly valued. It also provides important information that can aid diagnosis and management, for instance whether surgery should be conservative or radical. Radiotherapy and chemotherapy may also be altered. For example, platinum-based therapies are not standard treatment for breast cancer but can have utility in *BRCA* carriers[84,85]. Conversely, temozolomide is unlikely to be of benefit and may actually promote neoplastic progression in *MSH6* mutation carriers[86,87]. Identifying an underlying CPG mutation also provides important prognostic information; survival is significantly better for *BRCA2* mutation-positive patients with ovarian cancer but significantly worse for *BRCA2* mutation-positive patients with prostate cancer[88,89]. The likelihood of recurrence, a new primary and/or a second malignancy can all be increased in CPG mutation carriers who require ongoing review and consideration of tailored surveillance and/or risk-reducing interventions. Management of non-cancer-associated problems can also be important, for example certain *WT1* mutations result in insidious renal dysfunction, which requires monitoring and early intervention.

Targeted therapies

There is intense activity in the development of tailored therapies for cancer and strategies targeting CPGs and their constituent pathways have been among the most innovative and fruitful. The rationale is to harness knowledge of the underlying cause of cancer to identify tumour-specific vulnerabilities that can be therapeutically exploited. The simplest model is in cancers caused by gain-of-function mutations, which can be directly downregulated by inhibitors such as imatinib (*KIT* and *PDGFRA*), vandetanib (*RET*) and foretinib (*MET*)[90–92]. Trying to switch on genes that have been mutationally inactivated is more challenging. A direct approach of using compounds that 'read-through' stop codons is showing some promise, as is gene therapy[93,94]. Inhibiting a pathway member that is upregulated as a result of the CPG mutation has also had success. For example, everolimus, an mTOR inhibitor, is now approved for treatment of astrocytomas in tuberous sclerosis, and vismodegib, which inhibits the hedgehog pathway, has been found to be beneficial for basal-cell nevus syndrome patients with *PTCH1* mutations[95,96]. Perhaps, the most innovative approach has been through inducing synthetic lethality. PARP inhibitors, which cause lethality in *BRCA* deficient tumour cells but not normal cells with monoallelic mutations exemplify this approach, which is now being pursued for other CPGs[97–99]. Currently, these therapies are

still largely being evaluated in research studies and clinical trials, but there is optimism that identification of CPG mutations will increasingly lead to personalized management for patients with cancer.

Screening and prevention

An important benefit of CPG testing is being able to provide information about cancer risks to relatives. One of the unusual characteristics of CPGs is their capacity to serve as a biomarker of future disease. Identifying a CPG mutation can provide a window of opportunity to implement surveillance and/or risk-reducing measures that mitigate or prevent cancer. Naturally, the type of screening is determined by the type of cancer but most often involves imaging to detect a lesion before it presents clinically. Sometimes, a biochemical marker of risk can be measured, such as catecholamines or calcitonin in individuals who are at risk of phaeochromocytoma or thyroid cancer, respectively[100]. The presumption is that if a cancer is detected early, treatment and survival will be improved, although this has rarely been proven and for some cancers the available evidence does not suggest a benefit[101]. Prevention usually involves surgical removal of the at-risk tissue and is necessarily reserved for non-essential organs in individuals at very high risk, such as the stomach in *CDH1* mutation carriers, the thyroid in *RET* mutation carriers and the colon in *APC* mutation carriers[100,102,103]. Chemoprevention is an attractive strategy, but to date there have been few applications. A notable exception is mismatch repair gene mutation carriers, in whom the risk of colorectal cancer is significantly reduced by daily aspirin[104]. Of equal value, although commanding much less fanfare, is the use of CPG mutation testing in identifying people who do not have a familial CPG mutation. Such individuals do not have the high cancer risk of their relatives, reducing anxiety for themselves and their offspring. They also typically do not require costly, intensive screening or preventative interventions.

Pitfalls in CPG research and clinical practice

The study of CPGs has led to tremendous scientific and medical advances of broad and lasting impact. However, the field has been hampered by incorrect interpretations of genetic data, which can have substantial negative consequences.

The first major problem is the incorrect classification of a gene as a CPG. There are, unfortunately, dozens of genes in widely-used databases such as OMIM (Online Mendelian Inheritance in Man) and HGMD (The Human Gene Mutation Database) that are designated CPGs but for which the evidence is at best uncertain. For most, interpretation of the available data strongly suggests the gene does not confer high or moderate risks of cancer. There are various reasons for these misclassifications. Until recently, the extent of human coding variation was poorly appreciated, leading to overestimation of the likely causal link between the presence of a gene variant and cancer in an individual. This problem is actually increasing with exome-sequencing studies, with many researchers failing to appreciate that rare coding variations, including putative deleterious mutations, are collectively common[105,106]. The most widespread misconception is that the absence from the control group of a specific rare mutation that was identified in cancer cases provides evidence of causality, whereas usually it merely provides additional evidence that it is rare.

Over-extrapolation of concepts and data is a pervasive problem in CPG research and manifests in many ways. First, it is often presumed that if a gene mutation causes one cancer that any other cancer that occurs in a mutation carrier is also probably attributable to that gene, whereas frequently it will be coincidental because cancer is very common. Second, it is frequently incorrectly assumed that because one mutation class (for example, truncating mutations) predisposes carriers to cancer that variants in other classes (for example, missense mutations) are also causative, but many will be rare, innocuous variants. Third, it is commonly assumed that if some genes in a pathway are CPGs then variants in other gene members of that pathway are de facto likely to predispose carriers to cancer. Fourth, it is widely thought that cancer risks of CPG mutations are constant and can be extrapolated from one context to another, whereas many factors can influence the clinical expression of a CPG mutation. Finally, it is often incorrectly assumed that if a variant is shown to have some kind of functional impact, this proves it is pathogenic. CPGs have multiple complex functions and the relationship between functional aberrations and clinical phenotype is typically unclear or unknown. There are very few CPG functional assays that have been validated as robust tests of clinical pathogenicity. Thus, although functional data can provide supportive evidence for pathogenicity it can very rarely serve as a substitute for robust genetic evidence.

The extent to which these presumptions lead to incorrect scientific inferences and inappropriate clinical management depends on the specific CPG and scenario. However, significant and unacceptable negative impacts can result, including unwarranted surgery in healthy individuals.

Future opportunities and challenges

The future for CPG research is very bright, both in terms of scientific discovery and clinical translation. Strong evidence from multiple sources indicates that more CPGs remain to be discovered. Exome and genome sequencing are ideally suited to their identification, although standards to ensure consistent, robust designation as a CPG are required. For familial and syndromic cancer conditions, exome sequencing methods developed for Mendelian disorders will probably be successful, and are already yielding new genes[17,18]. As with other common, complex conditions, identification of non-syndromic genes will remain challenging, at least until it is possible to sequence, analyse and interpret data from many thousands of individuals. Innovative sample and analytical prioritization strategies, in the spirit of those used so successfully in the past will thus probably have high utility over the next few years[106].

It is also important to recognize that in this Review I have focused on germline gene mutations with high or moderate risks of cancer. Other components of the genetic architecture of cancer predisposition, such as common variants with small effects are also important and the interplay of different genes and variants is a topical field that will probably reveal new and clinically relevant insights[78,81,107]. It is also increasingly apparent that many other mechanisms are likely to play a part. One emerging area is the role of mosaic mutations, particularly in individuals with multiple cancers. Genetic and epigenetic cancer-predisposing post-zygotic events have been identified, for example mosaic *HIF2A* mutations in individuals with paraganglioma and *H19* hypermethylation in children with bilateral Wilms tumour[108,109]. More recently, mosaic *PPM1D* mutations associated with increased risk of ovarian and breast cancer have been reported, although the mechanism of cancer association is currently unclear[110]. Even though a considerable proportion of genetic predisposition to cancer probably resides in CPGs, the genetic architecture of cancer predisposition includes other components, many of which may be undiscovered.

Opportunities to use CPGs to improve management of patients should be vigorously pursued. This will lead to optimized, personalized care for mutation carriers and will probably provide insights of broader relevance to cancer, as exemplified by countless CPG-based discoveries of the past. The rarity of CPG mutations impedes research, and improved networks and registries of mutation carriers would greatly enhance the field. Routine integration of germline CPG testing into clinical trials will be invaluable, as will better collaborative links between somatic and germline cancer research. Probably the most important goal, which would facilitate all of the points mentioned, is to increase availability of CPG testing to patients with cancer. Next-generation sequencing makes large-scale, high-throughput CPG testing possible and affordable, but the clinical infrastructure needed to appropriately deliver such testing requires development. Various initiatives are seeking to achieve this, such as the UK Mainstreaming Cancer Genetics programme (http://www.mcgprogramme.com).

It is imperative that comprehensive evaluation of known CPGs in large patient and population series is performed so their cancer risks, clinical phenotypes, genotype–phenotype associations, genetic and non-genetic modifying factors, and contribution to cancer can be clarified. Large-scale, international, integrated molecular and clinical databases and analyses will greatly facilitate these endeavours. Enthusiasm for feedback of

incidental findings is best tempered until these data are available. Recently, the American College of Medical Genetics issued a policy statement recommending that incidental findings in 24 CPGs should be returned to the patient, irrespective of age or specific consent[111]. This has stimulated intense debate about possible ethical and legal ramifications. However, scant attention has been given to the arguably more pressing concern of our insufficient knowledge about the clinical consequences of mutations identified opportunistically. As already discussed, there is evidence to suggest that the impact of incidental mutations may differ substantially from that of mutations detected in individuals with a clinical phenotype and thus more curation and more caution are required.

That being said, the impact of CPG mutations in the general population is of high interest and has the potential to provide health benefits and opportunities for cancer prevention. It is often assumed cancer surveillance is of intrinsic value and should automatically be instituted in CPG mutation carriers. However, for most surveillance programmes there is little or no actual evidence of an improvement in outcome. The lack of proven efficacy and the potential risks of screening, such as overdiagnosis, misdiagnosis, false positives and false negatives, are rarely discussed. The low frequency of CPG conditions makes randomized clinical surveillance trials challenging. It also leads to the misguided impression that ad hoc screening in individual CPG families is a trivial burden. In fact, instituting decades of surveillance to relatives in a single family can be a very considerable financial outlay. To implement this at the population level, which may insidiously occur as genome sequencing becomes routine, could spiral into sizeable strains on the capacity and purse of health services. This is particularly likely if individuals with rare variants of unproven pathogenicity are (inappropriately) included in enhanced surveillance programmes, as is currently often the case. To ensure consistent, appropriate and affordable management of at-risk individuals, there needs to be a grass-roots move away from reflex interventions and to application and adherence to the accepted criteria of effective screening tests[112]. In parallel, we need to invest energy in developing carefully considered and evaluated strategies that maximize the benefits of identifying people who are at increased risk of cancer. ■

Received 31 October; accepted 21 November 2013.

1. Broca, P. *Traite des tumeurs*. (Asselin, 1866).
Broca describes the strong family history of breast cancer in his wife's relatives and, controversially for the time, proposes that it is due to hereditary factors.
2. Boveri, T. *Zur Frage der Entstehung Maligner Tumoren*. (Gustav Fischer, 1914).
Boveri's seminal work proposed that genomic dysregulation is central to cancer and may be inherited in some circumstances.
3. Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68,** 820–823 (1971).
A statistical study of retinoblastoma predicted it was due to two mutational events, one of which was inherited in familial and bilateral cases.
4. Fung, Y. K. *et al*. Structural evidence for the authenticity of the human retinoblastoma gene. *Science* **236,** 1657–1661 (1987).
5. Varghese, J. S. & Easton, D. F. Genome-wide association studies in common cancers–what have we learnt? *Curr. Opin. Genet. Dev.* **20,** 201–209 (2010).
6. Chang, C. Q. *et al*. A systemic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur. J. Hum. Genet.* http://dx.doi.org/10.1038/ejhg.2013.161 (2013).
7. Stadler, Z. K., Gallagher, D. J., Thom, P. & Offit, K. Genome-wide association studies of cancer: principles and potential utility. *Oncology* **24,** 629–637 (2010).
8. Zuo, L. *et al*. Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma. *Nature Genet.* **12,** 97–99 (1996).
9. Nichols, A. F., Ong, P. & Linn, S. Mutations specific to the xeroderma pigmentosum group E Ddb$^-$ phenotype. *J. Biol. Chem.* **271,** 24317–24320 (1996).
10. Sijbers, A. M. *et al*. Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. *Cell* **86,** 811–822 (1996).
11. Stickens, D. *et al*. The *EXT2* multiple exostoses gene defines a family of putative tumour suppressor genes. *Nature Genet.* **14,** 25–32 (1996).
12. Pilia, G. *et al*. Mutations in *GPC3*, a glypican gene, cause the Simpson-Golabi-Behmel overgrowth syndrome. *Nature Genet.* **12,** 241–247 (1996).
13. Feder, J. N. *et al*. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.* **13,** 399–408 (1996).
14. Whitcomb, D. C. *et al*. Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature Genet.* **14,** 141–145 (1996).
15. Yu, C. E. *et al*. Positional cloning of the Werner's syndrome gene. *Science* **272,** 258–262 (1996).
16. Johnson, R. L. *et al*. Human homolog of patched, a candidate gene for the basal cell nevus syndrome. *Science* **272,** 1668–1671 (1996).
17. Comino-Méndez, I. *et al*. Exome sequencing identifies *MAX* mutations as a cause of hereditary pheochromocytoma. *Nature Genet.* **43,** 663–667 (2011).
This article reports the first CPG to be identified through exome sequencing.
18. Smith, M. J. *et al*. Loss-of-function mutations in *SMARCE1* cause an inherited disorder of multiple spinal meningiomas. *Nature Genet.* **45,** 295–298 (2013).
19. Hanks, S. *et al*. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in *BUB1B*. *Nature Genet.* **36,** 1159–1161 (2004).
20. Armanios, M. *et al*. Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc. Natl Acad. Sci. USA* **102,** 15960–15964 (2005).
21. Nicolaides, N. C. *et al*. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* **371,** 75–80 (1994).
22. Miyaki, M. *et al*. Germline mutation of *MSH6* as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genet.* **17,** 271–272 (1997).
23. Niemann, S. & Muller, U. Mutations in *SDHC* cause autosomal dominant paraganglioma, type 3. *Nature Genet.* **26,** 268–270 (2000).
24. Meijers-Heijboer, H. *et al*. Low-penetrance susceptibility to breast cancer due to *CHEK2**1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nature Genet.* **31,** 55–59 (2002).
This article reports the first clear example of a moderate penetrance hereditary CPG.
25. Seal, S. *et al*. Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nature Genet.* **38,** 1239–1241 (2006).
26. Rahman, N. *et al*. *PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature Genet.* **39,** 165–167 (2007).
27. Hao, H. X. *et al*. *SDH5*, a gene required for flavination of succinate dehydrogenase, is mutated in paraganglioma. *Science* **325,** 1139–1142 (2009).
28. Loveday, C. *et al*. Germline mutations in *RAD51D* confer susceptibility to ovarian cancer. *Nature Genet.* **43,** 879–882 (2011).
29. Lohmann, D. R. *RB1* gene mutations in retinoblastoma. *Hum. Mutat.* **14,** 283–288 (1999).
30. Malkin, D. *et al*. Germ line *p53* mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250,** 1233–1238 (1990).
In this study, the authors used a candidate gene approach to demonstrate that germline *TP53* mutations confer an increased risk of multiple cancers, often referred to as Li-Fraumeni syndrome.
31. Huff, V. *et al*. Evidence for *WT1* as a Wilms tumor (WT) gene: intragenic germinal deletion in bilateral WT. *Am. J. Hum. Genet.* **48,** 997–1003 (1991).
32. Nishida, T. *et al*. Familial gastrointestinal stromal tumours with germline mutation of the *KIT* gene. *Nature Genet.* **19,** 323–324 (1998).
33. Sévenet, N. *et al*. Constitutional mutations of the *hSNF5/INI1* gene predispose to a variety of cancers. *Am. J. Hum. Genet.* **65,** 1342–1348 (1999).
34. Taylor, M. D. *et al*. Mutations in *SUFU* predispose to medulloblastoma. *Nature Genet.* **31,** 306–310 (2002).
35. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of *CEBPA* in familial acute myeloid leukemia. *N. Engl. J. Med.* **351,** 2403–2407 (2004).
36. Chompret, A. *et al*. *PDGFRA* germline mutation in a family with multiple cases of gastrointestinal stromal tumor. *Gastroenterology* **126,** 318–321 (2004).
37. Bell, D. W. *et al*. Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. *Nature Genet.* **37,** 1315–1316 (2005).
38. Niemeyer, C. M. *et al*. Germline *CBL* mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nature Genet.* **42,** 794–800 (2010).
39. Wiesner, T. *et al*. Germline mutations in *BAP1* predispose to melanocytic tumors. *Nature Genet.* **43,** 1018–1021 (2011).
40. Al-Tassan, N. *et al*. Inherited variants of *MYH* associated with somatic G:CT:A mutations in colorectal tumors. *Nature Genet.* **30,** 227–232 (2002).
In this innovative approach, the mutational signature in the tumours was used to identify the underlying CPG.
41. Forbes, S. A. *et al*. *The Catalogue of Somatic Mutations in Cancer (COSMIC)* (Wiley, 2008).
The COSMIC database is a catalogue of somatic mutations that have been identified in cancer and has proved highly useful for many aspects of research.
42. Hudson, T. J. *et al*. International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
43. Hindorff, L. A. *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
44. Barrett, J. H. *et al*. Genome-wide association study identifies three new melanoma susceptibility loci. *Nature Genet.* **43,** 1108–1113 (2011).
45. Gao, L. B. *et al*. The association between *ATM* D1853N polymorphism and breast cancer susceptibility: a meta-analysis. *J. Exp. Clin. Cancer Res.* **29,** 117 (2010).
46. Stacey, S. N. *et al*. A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. *Nature Genet.* **43,** 1098–1103 (2011).
47. Rafnar, T. *et al*. Sequence variants at the *TERT-CLPTM1L* locus associate with many cancer types. *Nature Genet.* **41,** 221–227 (2009).
48. Nelson, N. D. & Bertuch, A. A. Dyskeratosis congenita as a disorder of telomere maintenance. *Mutat. Res.* **730,** 43–51 (2012).
49. Mocellin, S. *et al*. Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis. *J. Natl Cancer Inst.* **104,** 840–854 (2012).
50. Teslovich, T. M. *et al*. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
51. Morris, A. P. *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genet.* **44,** 981–990 (2012).

52. Rahman, N. & Scott, R. H. Cancer genes associated with phenotypes in monoallelic and biallelic mutation carriers: new lessons from old players. *Hum. Mol. Genet.* **16,** R60–R66 (2007).
53. Dixit, A. *et al*. Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS ONE* **4,** e7485 (2009).
54. Huff, V. Wilms' tumours: about tumour suppressor genes, an oncogene and a chameleon gene. *Nature Rev. Cancer* **11,** 111–121 (2011).
55. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour suppression. *Nature* **476,** 163–169 (2011).
56. Villanueva, A., Newell, P. & Hoshida, Y. Inherited hepatocellular carcinoma. *Best Pract. Res. Clin. Gastroenterol.* **24,** 725–734 (2010).
57. Rutter, J., Winge, D. R. & Schiffman, J. D. Succinate dehydrogenase — assembly, regulation and role in human disease. *Mitochondrion* **10,** 393–401 (2010).
58. Santen, G. W., Kriek, M. & van Attikum, H. SWI/SNF complex in disorder: SWItching from malignancies to intellectual disability. *Epigenetics* **7,** 1219–1224 (2012).
59. Sheppard, K., Kinross, K. M., Solomon, B., Pearson, R. B. & Phillips, W. A. Targeting PI3 kinase/AKT/mTOR signaling in cancer. *Crit. Rev. Oncog.* **17,** 69–95 (2012).
60. Slade, I. *et al*. DICER1 syndrome: clarifying the diagnosis, clinical features and management implications of a pleiotropic tumour predisposition syndrome. *J. Med. Genet.* **48,** 273–278 (2011).
61. Scott, R. H., Stiller, C. A., Walker, L. & Rahman, N. Syndromes and constitutional chromosomal abnormalities associated with Wilms tumour. *J. Med. Genet.* **43,** 705–715 (2006).
62. Scott, R. H. *et al*. Constitutional 11p15 abnormalities, including heritable imprinting center mutations, cause nonsyndromic Wilms tumor. *Nature Genet.* **40,** 1329–1334 (2008).
63. Gayther, S. A. & Pharoah, P. D. The inherited genetics of ovarian and endometrial cancer. *Curr. Opin. Genet. Dev.* **20,** 231–238 (2010).
64. Pacini, F., Castagna, M. G., Cipri, C. & Schlumberger, M. Medullary thyroid carcinoma. *Clin. Oncol.* **22,** 475–485 (2010).
65. Jafri, M. & Maher, E. R. The genetics of phaeochromocytoma: using clinical features to guide genetic testing. *Eur. J. Endocrinol.* **166,** 151–158 (2012).
66. Mavaddat, N. *et al*. Pathology of breast and ovarian cancers among *BRCA1* and *BRCA2* mutation carriers: results from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA). *Cancer Epidemiol. Biomarkers Prev.* **21,** 134–147 (2012).
67. Benusiglio, P. R. *et al*. *CDH1* germline mutations and the hereditary diffuse gastric and lobular breast cancer syndrome: a multicentre study. *J. Med. Genet.* **50,** 486–489 (2013).
68. Rausch, T. *et al*. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148,** 59–71 (2012).
69. Alexandrov, L. B. *et al*. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
70. Aoki, Y. *et al*. Germline mutations in *HRAS* proto-oncogene cause Costello syndrome. *Nature Genet.* **37,** 1038–1040 (2005).
71. Hafner, C. & Groesser, L. Mosaic RASopathies. *Cell Cycle* **12,** 43–50 (2013).
72. Horn, S. *et al*. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339,** 959–961 (2013).
This provides one of the clearest examples of specific promotor mutations that predispose to cancer, notably melanoma is not one of the carriers prominent in dyskeratosis congenita caused by exonic *TERT* mutations.
73. Goudie, D. R. *et al*. Multiple self-healing squamous epithelioma is caused by a disease-specific spectrum of mutations in *TGFBR1*. *Nature Genet.* **43,** 365–369 (2011).
74. Breast Cancer Linkage Consortium. Cancer risks in *BRCA2* mutation carriers. *J. Natl Cancer Inst.* **91,** 1310–1316 (1999).
75. Thompson, D. & Easton, D. Variation in cancer risks, by mutation position, in *BRCA2* mutation carriers. *Am. J. Hum. Genet.* **68,** 410–419 (2001).
76. Ford, D. *et al*. Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am. J. Hum. Genet.* **62,** 676–689 (1998).
77. Antoniou, A. *et al*. Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72,** 1117–1130 (2003).
This is the largest analysis of cancer risks in CPG mutation carriers, demonstrating that the average risks in relatives of cancer cases unselected for family history is lower than in those with a family history of the disease.
78. Antoniou, A. C. *et al*. Common breast cancer-predisposition alleles are associated with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers. *Am. J. Hum. Genet.* **82,** 937–948 (2008).
79. Couch, F. J. *et al*. Genome-wide association study in *BRCA1* mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.* **9,** e1003212 (2013).
80. Gaudet, M. M. *et al*. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. *PLoS Genet.* **6,** e1001183 (2010).
81. Moorman, P. G. *et al*. Evaluation of established breast cancer risk factors as modifiers of BRCA1 or BRCA2: a multi-center case-only analysis. *Breast Cancer Res. Treat.* **124,** 441–451 (2010).
82. Chompret, A. *et al*. *P53* germline mutations in childhood cancers and cancer risk for carrier individuals. *Br. J. Cancer* **82,** 1932–1937 (2000).
83. Figueiredo, B. C. *et al*. Penetrance of adrenocortical tumours associated with the germline *TP53* R337H mutation. *J. Med. Genet.* **43,** 91–96 (2006).
84. Byrski, T. *et al*. Results of a phase II open-label, non-randomized trial of cisplatin chemotherapy in patients with *BRCA1*-positive metastatic breast cancer. *Breast Cancer Res.* **14,** R110 (2012).
85. Turner, N. C. & Tutt, A. N. Platinum chemotherapy for *BRCA1*-related breast cancer: do we need more evidence? *Breast Cancer Res.* **14,** 115 (2012).
86. Hunter, C. *et al*. A hypermutation phenotype and somatic *MSH6* mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66,** 3987–3991 (2006).
87. Scott, R. H. *et al*. Medulloblastoma, acute myelocytic leukemia and colonic carcinomas in a child with biallelic *MSH6* mutations. *Nature Clin. Pract. Oncol.* **4,** 130–134 (2007).
88. Vencken, P. M. *et al*. Outcome of *BRCA1*- compared with *BRCA2*-associated ovarian cancer: a nationwide study in the Netherlands. *Ann. Oncol.* **24,** 2036–2042 (2013).
89. Castro, E. *et al*. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. *J. Clin. Oncol.* **31,** 1748–1757 (2013).
90. Bachet, J.-B. *et al*. Diagnosis, prognosis and treatment of patients with gastrointestinal stromal tumour (GIST) and germline mutation of *KIT* exon 13. *Eur. J. Cancer* **49,** 2531–2541 (2013).
91. Logan, T. F. Foretinib (XL880): c-MET inhibitor with activity in papillary renal cell cancer. *Curr. Oncol. Rep.* **15,** 83–90 (2013).
92. Wells, S. A. Jr *et al*. Vandetanib for the treatment of patients with locally advanced or metastatic hereditary medullary thyroid cancer. *J. Clin. Oncol.* **28,** 767–772 (2010).
93. Bordeira-Carriço, R., Pego, A. P., Santos, M. & Oliveira, C. Cancer syndromes and therapy by stop-codon readthrough. *Trends Mol. Med.* **18,** 667–678 (2012).
94. Aiuti, A. *et al*. Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott–Aldrich syndrome. *Science* **341,** 1233151 (2013).
In this study, a lentiviral vector encoding functional WASP was used to genetically correct haematopoeitic stem cells, which were reinfused into three patients with Wiskott–Aldrich syndrome, with improved clinical symptoms.
95. Józwiak, S., Stein, K. & Kotulska, K. Everolimus (RAD001): first systemic treatment for subependymal giant cell astrocytoma associated with tuberous sclerosis complex. *Future Oncol.* **8,** 1515–1523 (2012).
96. Tang, J. Y. *et al*. Inhibiting the hedgehog pathway in patients with the basal-cell nevus syndrome. *N. Engl. J. Med.* **366,** 2180–2188 (2012).
97. Farmer, H. *et al*. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434,** 917–921 (2005).
A synthetic lethality strategy was utilised in this study to therapeutically target the DNA repair defect in BRCA deficient cells.
98. Fong, P. C. *et al*. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361,** 123–134 (2009).
99. Brough, R., Frankum, J. R., Costa-Cabral, S., Lord, C. J. & Ashworth, A. Searching for synthetic lethality in cancer. *Curr. Opin. Genet. Dev.* **21,** 34–41 (2011).
100. Wells, S. A. Jr, Pacini, F., Robinson, B. G. & Santoro, M. Multiple endocrine neoplasia type 2 and familial medullary thyroid carcinoma: an update. *J. Clin. Endocrinol. Metab.* **98,** 3149–3164 (2013).
101. Reade, C. J., Riva, J. J., Busse, J. W., Goldsmith, C. H. & Elit, L. Risks and benefits of screening asymptomatic women for ovarian cancer: a systematic review and meta-analysis. *Gynecol. Oncol.* **130,** 674–681 (2013).
102. Rozen, P. & Macrae, F. Familial adenomatous polyposis: the practical applications of clinical and molecular screening. *Fam. Cancer* **5,** 227–235 (2006).
103. Seevaratnam, R. *et al*. A systematic review of the indications for genetic testing and prophylactic gastrectomy among patients with hereditary diffuse gastric cancer. *Gastric Cancer* **15** (Suppl 1), 153–163 (2012).
104. Burn, J., Mathers, J. C. & Bishop, D. T. Chemoprevention in Lynch syndrome. *Fam. Cancer* **12,** 707–718 (2013).
105. Abecasis, G. R. *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
106. Snape, K. *et al*. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res. Treat.* **134,** 429–433 (2012).
107. Turnbull, C. *et al*. Gene–gene interactions in breast cancer susceptibility. *Hum. Mol. Genet.* **21,** 958–962 (2012).
108. Zhuang, Z. *et al*. Somatic *HIF2A* gain-of-function mutations in paraganglioma with polycythemia. *N. Engl. J. Med.* **367,** 922–930 (2012).
Somatic gain-of-function mutations in *HIF2A* predispose carriers to certain tumours, including multiple tumours within an individual, but are not hereditary.
109. Scott, R. H. *et al*. Stratification of Wilms tumor by genetic and epigenetic analysis. *Oncotarget* **3,** 327–335 (2012).
110. Ruark, E. *et al*. Mosaic *PPM1D* mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493,** 406–410 (2013).
111. Green, R. C. *et al*. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15,** 565–574 (2013).
112. Wilson, J. M. G. & Jungner, G. *Principles and Practice of Screening for Disease* (WHO, 1968).

Supplementary Information is linked to the online version of the paper at go.nature.com/i9obmq

Acknowledgements I am very grateful to many colleagues with whom I have discussed discovery, characterization and clinical translation of CPGs over the past 15 years in particular M. Stratton, H. Hanson and C. Turnbull. I am indebted to A. Strydom for editorial assistance, S. Hanks for construction of Fig. 1 and S. Mahamdallie, B. De Souza, C. Turnbull and E. Ruark for input into the Supplementary Information.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/i9obmq. Correspondence should be addressed to N.R. (nazneen.rahman@icr.ac.uk).

REVIEW

doi:10.1038/nature12982

Circuit dynamics of adaptive and maladaptive behaviour

Karl Deisseroth[1,2,3]

The recent development of technologies for investigating specific components of intact biological systems has allowed elucidation of the neural circuitry underlying adaptive and maladaptive behaviours. Investigators are now able to observe and control, with high spatio–temporal resolution, structurally defined intact pathways along which electrical activity flows during and after the performance of complex behaviours. These investigations have revealed that control of projection-specific dynamics is well suited to modulating behavioural patterns that are relevant to a broad range of psychiatric diseases. Structural dynamics principles have emerged to provide diverse, unexpected and causal insights into the operation of intact and diseased nervous systems, linking form and function in the brain.

A long-sought goal in psychiatry is to understand the concrete mechanistic distinctions between adaptive, physiological behavioural states and psychiatric-disease-related behavioural states, which are symptomatically complex in a way that can seem inaccessible to physical interpretation or precise intervention. This situation stands in contrast to, for example, our mechanistic understanding of the equally diverse symptoms of heart failure (ranging from shortness of breath to swelling of extremities), which can be concretely explained and understood using a single physical formulation with robust causal explanatory power: reduced pumping action of the heart muscle. Unfortunately, such causal tissue-level understanding does not yet exist in psychiatry[1–3], despite pioneering genetic and pharmacological studies that have provided causal insight at the molecular level, and brain-wide observational techniques such as BOLD–fMRI (blood-oxygen-level-dependent functional–magnetic resonance imaging) and electroencephalography that have provided circuit-level correlates (but without causal information)[1,2,4].

Although neural connections are presumed to transmit information-rich signalling streams, long-range projections might be relevant to neuropsychiatric symptoms through neural dynamics (how brain activity is managed and coordinated over time and space), as significantly as through circuit coding (how detailed neural information is actually represented). Recent advances in neuroscience have revealed that surprisingly potent and specific control of complex behaviours arises from modulating circuit projection dynamics, even though we have very little causal knowledge of information coding or representation. This Review will consider the technological advances that allow this insight into the causal underpinnings of maladaptive behaviour, in the context of the structural dynamics findings that have been made accessible by the application of new technologies.

Clinical context

Deep-brain stimulation (DBS) with microelectrodes has set the stage for real-time causal intervention in neuropsychiatric conditions. Microelectrodes can be placed in anatomically well-defined locations for region-specific neuromodulation[5], and have delivered potent and specific elicitation or resolution of subsets of symptoms in disease states ranging from parkinsonism to depression[6–11]. However, DBS mechanisms and effects have been mysterious, partly because DBS electrode stimulation directly causes mixed patterns of excitation and inhibition in diverse local cells and in axons from distant sources that are passing through and may be unrelated to the function of the implanted region[5]. An interesting pattern is emerging from our clinical experience of psychiatric DBS; many of the most promising targets involve electrode contact placement in white matter (the long-range axonal connections that wire brain regions together) rather than in grey matter or cell-body-dense regions[6,8]. DBS targets that fit this pattern include subcallosal cingulate white matter for depression, the medial forebrain bundle for depression and anhedonia (the lack of enjoyment of normally rewarding experiences), the anterior commissure for depression, zona incerta (ZI) or subthalamic nucleus (STN)-proximal white matter tracts for depression and obsessive–compulsive disorder (OCD), ventral capsule white matter in OCD and depression, inferior thalamic peduncle white matter in OCD and depression, and even the vagus nerve axon bundle (including afferent fibres to the solitary tract nucleus) in depression[6]. However, the mechanistic implications of this pattern are unclear. The brief high-frequency pulses that are typically delivered through DBS electrodes may simply be better suited to stimulating axons than cell bodies[5,12], and as point sources, DBS electrodes will more efficiently control a tract of axons in a relatively small volume than directly control a larger region of cell bodies in grey matter[13].

Apart from these important technical considerations, long-range projection properties (whether adaptive or maladaptive) may represent suitable final common pathways for governing psychiatry-related behavioural states. Although structurally defined, these features of brain anatomy can place natural bounds on tissue dynamics, and can do so independently of neural coding per se. The excitability, myelination and conduction properties, valence (excitation, inhibition or modulation) and net synaptic strength of axonal connections that constitute a particular long-range projection in the brain (all physical quantities that can be set by genetics, development and plasticity) are suited to govern circuit-level dynamical properties that have previously been linked to psychiatric disease, such as excitation/inhibition (E/I) balance[14–17], synchronization of activity across and within brain regions[18–25], and extent of activity propagation through brain regions[26–28]. Until recently, testing the causal impact of specific long-range projections on behaviour was not possible — achievable only with a new experimental methodology.

[1]Department of Bioengineering, Stanford University, Stanford, California 94305, USA. [2]Department of Psychiatry, Stanford University, Stanford, California 94305, USA. [3]Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA.

Technology

Stimulated, in part, by this clinical context, technologies that work with specificity at the level of circuit wiring in animal models have become the focus of much recent research interest in neuroscience. New approaches that allow interrogation of specific nervous system connections include projection-defined activity control, as well as projection activity mapping and structural mapping. The outcome of this technology development has been the ability to control and observe specific connections within the intact nervous system through precise circuit-level measurements and interventions.

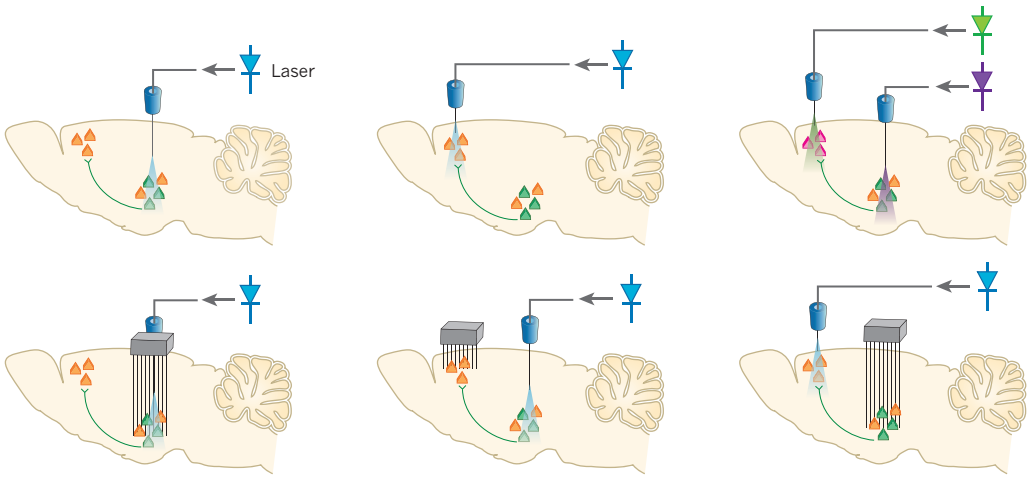
Achieving control over cells defined by specific connectivity is difficult, but essential to progress along the path towards the causal understanding of function and dysfunction of neural circuitry. Already, research in some invertebrate nervous systems (within which defined cells can be selectively controlled by microelectrodes) has demonstrated the importance of knowing the wiring patterns of the cells targeted for control, to achieve predictable influence over the complex and sensitive dynamical properties of even small networks of neurons[29]. However, in mammalian systems, until very recently it was not possible to control the high-speed dynamics of cells defined by wiring, because even when directed to fibre bundles or white matter, electrodes cannot select as the initial direct target a particular projection defined by cell-body origin and trajectory. Optogenetics[30,31] (Box 1), which allows researchers to have direct control over cells defined by projection pattern[4,32–34], was initially applied to brain disease in a study probing DBS mechanisms[13]. It was found that amelioration of parkinsonian symptoms could be most robustly achieved when the light-sensitive elements were chiefly afferent axons rather than local cell bodies[13,35]. Shortly after, optogenetic control of specific trajectory-defined projections in behaviour was achieved[36], unexpectedly revealing endogenous inhibition of anxiety-related states in the amygdala, in a study using gain- and loss-of-function interventions to targeted amygdala projections. Optogenetic approaches have since allowed projection-defined activity control in behaviours related to reward, motivation, depression, social interaction, compulsions, cognition and other domains of normal and maladaptive brain function (discussed later). Of note, projection targeting is only one application of optogenetics that is distinct from other approaches such as controlling distinct cell types within a region to assess the impact on physiology or behaviour, developing prosthetic or repair strategies, mapping detailed wiring patterns, and perturbing dynamics precisely during high-dimensional observation (recording or imaging) for population and state space analysis.

Controlling activity in cells defined by projection pattern would ideally be complemented by the observation of native activity and structure in these cells (Box 1). Indicators of neural activity could in principle be used to selectively record the activity of a projection *in vivo*, for example with genetically encoded calcium indicators (GECIs)[37], which can be introduced through focal injection of increasingly sophisticated targeting viral vectors[24,38–41]. Cells could be connectivity-defined for activity imaging if this transduction process were to occur in a wiring-dependent manner,

BOX 1

Typical projection-targeting experimental configurations

Opsin genes carried by viral vectors are focally injected into the brain. Top left, a projecting population has been defined by injection of a retrograde viral tool to deliver an opsin or recombinase[48,124] into the upper left cell population; to provide projection specificity of the control, light is delivered by a fibre-optic not to the same location but to a distinct focal region of upstream cells (typically about 1 mm^3 in size). Top middle, in a technically different approach[4,32–34,125], microbial opsins — ideally enhanced with membrane-trafficking motifs from mammalian channels[124,126] — are transported down axons, which become the photosensitive elements in downstream fibre-optic-targeted structures. In this case, a well-tolerated (chiefly non-retrograde) adeno-associated virus or lentiviral vector may be used, but weeks are needed for effective opsin transport. Inhibition of the projection can be achieved with an inhibitory opsin gene[36,92], encoding a chloride or proton pump, whereas stimulation of cells defined by the projection can be achieved with an excitatory opsin gene[34,123], encoding a cation channel (in the case of the latter, by using channelrhodopsins, antidromic action potential backpropagation from the site of light delivery may occur, which can confer the feature of fully recruiting cells defined as possessing the specified projection among other possible projections). Top right, recording with various modalities (activity photometry[122] is shown here with an additional readout optical fibre, or electrical recording as in bottom row) at any point of interest in the brain allows assessment of the circuit dynamical underpinnings of the behavioural changes observed or photo-tagging of cells defined by projection. Bottom left, high-speed multi-unit electrical recording; spikes corresponding to cells defined by genetically determined opsin expression are identified through sufficiently low-latency spikes after a light pulse. Bottom right, same as at left except cells are defined for photo-tagging by projection illumination rather than projection transduction. Adapted with permission from ref. 127.



and certain rabies- or herpes-virus-based tools (although these are limited in utility by toxicity to target cells) will transduce axon terminals at the projection location[42–45], in principle allowing the observation of activity in projection-defined cells. Although these upstream cells may be distributed over large distances, activity imaging is carried out focally, thereby spatially defining cells that give rise to a specific projection (Box 1).

Activity recording of projection-defined cells (although not of the projection itself) when animals are exhibiting a certain behaviour has been achieved using optogenetics itself as an identification tool to discriminate recorded cells[46–48] based on wiring phenotype. Through this approach, an excitatory opsin is conditionally expressed based on projection pattern by the introduction of the opsin-carrying axon-transducing virus (for example, rabies or herpes) into a terminal field of the projection. Then pulses of light delivered to the cell-body location may elicit the lowest-latency (directly excited) electrical spikes from cells bearing the projection of interest (Box 1); the characteristic waveforms of these fast-responding cells can be quantitatively defined so that subsequent multielectrode recording during behaviour will allow the recognition of spikes from these projection-defined cells[46–48]. Although the specificity of this 'tagging' strategy is not absolute unless spike latencies in the low-ms range can be achieved or (more likely) details of the local circuitry preclude fast indirect synaptic activation[33], this combination of optogenetics and multielectrode recording has allowed, in some settings, the real-time activity recording of cells defined by the projection target (which need not be identical to the distant activity of the projection termination itself) during mammalian behaviour.

Finally, to complement projection-based activity control and projection-based activity measurement, detailed observation of the brain-wide physical form of the projections in question will be important. Visualization of neural projections linked to causality information and molecular descriptors will deepen our understanding of the neural structural dynamics that underlie behaviour, although so far pioneering sectioning or ablative electron microscopy and array tomography methods[49–53] (in some cases with correlational information on activity patterns) are not readily linkable with causal information on the behavioural significance of the connections. Clearing chemicals such as Sca*l*e[54], BABB[55,56], SeeDB[57] and ClearT[58] allow degrees of brain transparency (although without molecular phenotyping because the resultant brain tissue remains largely impermeable to macromolecular labels such as antibodies). High-throughput projection-mapping[50,59] tools involve thin-sectioning and reconstruction by alignment, and, in an approach without sectioning, the electrochemical technology CLARITY allows the construction of crosslinked hydrogels from within tissue and subsequent electrophoretic removal of membrane lipids, thereby allowing the penetration of photons and macromolecular labels throughout the intact mammalian brain[60,61] (Fig. 1). All of these approaches can be linked to causal or activity information on the projections observed if registration of data sets is conducted; projection activity could be accessible not only with acute slice approaches[62] but also with *in vivo* behavioural methods because opsin or GECI probes can be integrated with, or include, cell-filling fluorescent labels that allow projection mapping by light microscopy after the completion of behavioural testing or activity imaging. If needed, light-microscopy-based projection maps can also be linked to ultrastructural information[60] to describe the synaptic properties and targets of the projection.

Circuit-level findings

Together, the technologies delineated have opened the door to determining and understanding the causal significance of defined projections in the control of behaviour. Here, I discuss behavioural findings that have emerged about the interrelationships between adaptive and maladaptive behaviour through experiments probing causal structural dynamics at cellular resolution and within the intact brain. For the purposes of emphasis on behavioural and clinical insight, I highlight the major psychiatric symptom domains relating to anxiety and depression. A complete summary is provided in Table 1.

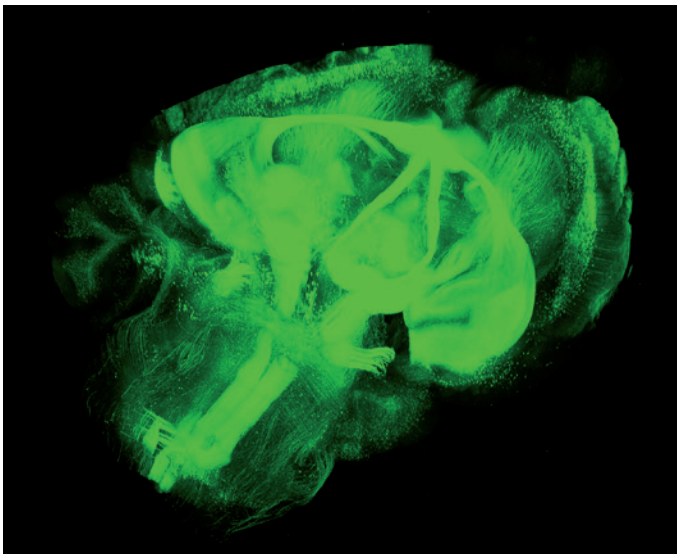


Figure 1 | Visualizing projections in the intact mammalian brain. The entire mouse brain has been processed with CLARITY[60,61] (with Thy-1 type projection neurons labelled in green); such brain wide structural data sets may be integrated with other data streams from the same animal, including activity records during behaviour (as obtained with GECIs), stable activity markers such as immediate-early genes, whole-mount molecular phenotyping, behavioural scores in the presence or absence of optogenetic or disease-model interventions, and opsin–fluorophore expression patterns. In general, whole-brain analysis is useful for interpretation; for example, use with multiple retrograde labels may help to determine whether individual cells include collaterals projecting to multiple sites, an important consideration in projection targeting. Adapted with permission from ref. 61.

Causal projection dynamics linked to anxiety

Anxiety disorders, which include generalized anxiety disorder, panic disorder, post-traumatic stress disorder and OCD, represent the most prevalent class of psychiatric condition. Building on earlier work that implicates structures such as the extended amygdala in fear and anxiety, and the striatum in OCD, recent studies have made headway in our understanding of the causal projection dynamics of anxiety-related symptoms. In an emerging theme, fundamentally different results were often obtained when projection-defined circuit elements were recruited, compared with the recruitment of region-defined cell-body populations.

In initial anxiety-related investigations[36], freely moving mice were assessed using the elevated-plus maze and the open-field test, which allow inference of apprehension in the absence of immediate threat. When the cell bodies of excitatory (CaMKIIα-expressing) neurons were optogenetically stimulated in the basolateral amygdala (BLA) (without regard to projection target), an immediate (and immediately reversible) anxiogenic effect was observed in which subjects began to avoid exposed areas of test arenas, without other changes in locomotor performance. Conversely, when these neurons were optogenetically inhibited, an anxiolytic effect was seen[36]; together, these results were consistent with classical anxiogenic models of the BLA.

However, more refined projection targeting revealed unanticipated anxiolytic circuitry embedded within this otherwise anxiogenic BLA circuit environment. When only a subpopulation of the CaMKIIα-expressing elements — namely lateral projections from the BLA towards the central nucleus of the amygdala (CeA) — were stimulated instead, a prominent anxiolytic effect was observed[36]. Conversely, inhibition of this projection elicited anxiogenesis[36]. This surprising arrangement (that required optogenetic projection targeting to resolve) is well-suited to providing diverse access nodes for swift and versatile endogenous modulation[2,63], and no doubt further anatomical and functional complexities will be elaborated. For example, studies of conditioned fear[64,65] have identified two functionally and genetically distinct types of unit in the centrolateral component of the CeA, and optogenetic stimulation of one of these

Table 1 | Projection-defined activity dynamics causally implicated in psychiatric-disease-related behaviour

Projection origin	Projection target	Elicited activity in projection	Behaviour	References
mPFC	LHb	Excitation	Passive/immobile (swim test)	76
	DRN	Excitation	Active/mobile (swim test)	76
		Inhibition	Passive/immobile (swim test)	76
	NAc	Excitation	Appetitive	93
BLA	CeL/CeA	Excitation	Anxiolytic (risk)	62
		Inhibition	Anxiogenic (risk)	62
	adBNST	Excitation	Anxiolytic (risk, respiratory)	68
		Inhibition	Anxiogenic (risk, respiratory)	68
	NAc	Excitation	Appetitive	92, 93
		Inhibition	Aversive	92
	vH	Excitation	Anxiogenic (risk)	121
		Inhibition	Anxiolytic (risk)	121
adBNST	LH	Excitation	Anxiolytic (risk)	68
	PB	Excitation	Anxiolytic (respiratory)	68
	VTA	Excitation	Appetitive	68
vBNST	VTA	Excitation (glutamatergic)	Aversive, anxiogenic	100
	VTA	Excitation (GABAergic)	Appetitive, anxiolytic	100
	LH	Excitation (GABAergic)	Increased food consumption	88
	LH	Inhibition (GABAergic)	Decreased food consumption	88
vH	NAc	Excitation	Appetitive	93
LDT	VTA	Excitation	Appetitive	97
LHb	VTA	Excitation	Aversive	97
EPN	LHb	Excitation	Aversive	99
VTA TH	NAc	Excitation	Appetitive	81
	mPFC	Excitation	Aversive	122
	LHb	Excitation (GABAergic-TH)	Appetitive	123
ARC AgRP	PVH	Excitation	Increased food consumption	87
vmOFC	VMS	Excitation (chronic: probably favouring potentiation of projection)	Increased grooming (wild-type mice)	70
LOFC	CMS	Excitation (acute: probably favouring inhibitory cell targets)	Reduced grooming (Sapap3-mutant OCD mouse model)	71

adBNST, anterodorsal bed nucleus of the stria terminalis; ARC AgRP, agouti-related peptide cells in arcuate nucleus; BLA, basolateral amygdala; CeA, central amygdala; CeL, centro-lateral amygdala; CMS, centromedial striatum; DRN, dorsal raphe nucleus; EPN, entopeduncular nucleus; LDT, laterodorsal tegmentum; LH, lateral hypothalamus; LHb, lateral habenula; LOFC, lateral orbitofrontal cortex; mPFC, medial prefrontal cortex; NAc, nucleus accumbens; PB, parabrachial nucleus; PVH, paraventricular hypothalamus; TH, tyrosine hydroxylase; vBNST, ventral BNST; vH, ventral hippocampus; vmOFC, ventromedial orbitofrontal cortex; VMS, ventromedial striatum; VTA, ventral tegmental area.

populations inhibited neurons projecting to periaqueductal grey[65]; this work illustrates that sufficiently precise genetic cell-type targeting (where feasible) can give rise to a kind of projection targeting, and together with earlier work has deepened our understanding of fear behaviour[66,67].

Although the initial anxiety study[36] focused on risk-avoidance behaviour (in the elevated-plus maze and the open-field test), the clinical anxious state is not characterized by risk-avoidance alone. Physiological phenomena such as respiratory-rate changes and aversive subjective sensations are an enormously important part of the clinical picture, contribute to morbidity and mortality and can be studied quantitatively in animal models. A subsequent paper[68] extended the projection targeting concept in anxiety to study assembly of the anxious state from multiple separable features. This study found that a long-range projection from the BLA to the bed nucleus of the stria terminalis (BNST) was unexpectedly able to favour anxiolysis, that within the BNST itself oppositional subnuclei favoured anxiogenesis or anxiolysis, and that distinct outgoing projections from the anxiolytic subnucleus (the anterodorsal BNST, adBNST) each recruited distinct features of anxiolysis[68] (Fig. 2). Surprisingly, the BNST projection to the lateral hypothalamus was found to recruit only a behavioural risk-avoidance feature, the projection to the ventral tegmental area (VTA) was found to recruit only an appetitive (positive-valence conditioning or rewarding) feature, and the projection to the brainstem parabrachial nucleus was found to recruit only a reduced respiratory-rate feature[68] (Fig. 2). Projection targeting allowed the determination of the causal nature of this feature separability, illustrating aspects of how complex behavioural states may be coordinately assembled and disassembled by projection dynamics. It is worth underscoring here that neural codes per se, in terms of the spiking representations of information, are not understood for any of these internal states or behavioural outputs, but relatively complex and specific behaviours can still be elicited and suppressed using projection targeting of circuit dynamics.

OCD highlights some of the challenges (and illustrates some of the solutions) for the study of disease-related symptoms in animal models. The disorder is characterized both by symptoms that are difficult to assess in experimental animals (recurring and intrusive thought patterns) and by symptoms with clear rodent correlates (compulsive behaviours that are poorly suppressed even when adverse consequences result). As with many psychiatric diseases, insight into the overall syndrome may be obtained in animal models by focusing on those symptom domains that can be recapitulated experimentally. For example, circuits involving the orbitofrontal cortex (OFC) and striatum have previously been correlated with clinical OCD[69], and two recent papers have now causally tested these circuits using projection targeting in mouse models of compulsive behaviour. One study[70] involved driving cells projecting from the ventromedial OFC to the ventromedial striatum (VMS), and tracked grooming behaviour (compulsive grooming bears some similarity to behaviours in the clinical condition). Repeated bursts of activation of this projection led to increased grooming over days, suggesting a mechanism by which pathological behavioural patterns subserved by that projection could become difficult to suppress. Another group began with a model of compulsive grooming (Sapap3 mutant mice) and targeted a projection between the lateral OFC and the centromedial striatum (CMS)[71]. Acutely driving this projection recruited feedforward inhibition in the CMS and gave rise to suppression of compulsive grooming. OFC is important for many behavioural and cognitive processes, including aspects of attention, reversal learning, risk responsiveness and

response inhibition[69,72]; together, these papers causally implicated specific projections from the OFC to the striatum in modulating compulsive grooming relevant to OCD.

Causal projection dynamics of depression-related behaviour

One of the core diagnostic criteria for major depressive disorder is the clinically defined symptom of hopelessness, a profound negativity about the future that can be behaviourally manifested as a predilection to discount the value of choices or effort in the present. Such a behavioural pattern can be pathological and give rise to severe morbidity (and mortality from suicide); however, when environmental conditions are adverse, a 'passive-coping' behavioural state — in which minimal energy is expended — could in fact be more adaptive than an active-coping state[63,73–76]. A recent report described the results of modulating dynamics in targeted projections arising from the prefrontal cortex, in the setting of a behavioural task (the automated forced swim test, FST) designed to detect temporally precise transitions between active and passive coping[76].

When cells with a projection between the medial prefrontal cortex (mPFC, an anterior forebrain structure involved in executive function and planning) and the dorsal raphe nucleus (DRN, a major locus of serotonergic neurons) were optogenetically stimulated during the automated FST, rats were found to shift towards an active-coping behavioural regime (swimming or climbing); the same fibre-optic-targeted intervention did not alter nonspecific locomotor activity assessed in the open field[76]. Surprisingly, the opposite finding was seen when cells defined by projections from the mPFC to the lateral habenula (LHb) were targeted; rats shifted toward the passive-coping (floating or immobile) regime, illustrating the importance in optogenetics (for true functional specificity) of not just cell-type targeting but also projection targeting allowed by the fibre-optic neural interface[40]. And when cells defined by projection from the mPFC to the BLA were optogenetically recruited, no behavioural effects in this test were seen[76]. It was also found that targeting corresponding focal cell-body regions (instead of defined projections) was not well-suited to elicitation of this specific class of behavioural effect. First, when DRN cells were driven optogenetically, but nonspecifically, increased active behaviour was seen in the FST but also in the open field, indicating elicitation of a less-specific behavioural state[76]. Second, when mPFC cells were generally recruited without specifying projection class, no transition between active and passive behavioural state was seen, probably partly because oppositional projections were recruited[76].

The converse of depression-related behavioural states (with maladaptively low value assigned to actions or available choices) would be states in which predictions are maladaptively positive. Manic and hypomanic behaviour may fall into this category in the form of the core criteria of increased goal-directed and risk-taking behaviour (along with certain character traits and behavioural patterns such as pathological gambling, especially those associated with dopamine (DA)-system modulation[77,78] in which perception of risk[79,80] is impaired and/or perceived likelihood of positive outcomes is inflated). In a transgenic rat line that allowed opsin expression only in tyrosine hydroxylase (TH)-expressing cells[81] in the VTA (that is, DA neurons), stimulation of these cells was recently found to favour the active-coping behavioural pattern in the automated FST without nonspecific locomotion effects in the open field[75]. Electrical recording in the anatomical target of one of these outgoing projections (the nucleus accumbens, NAc) from multiple single units during the automated FST revealed that modulating DA neuron dynamics altered key aspects of the neural representation of action in the NAc[75]. This particular projection (VTA–NAc) may be clinically relevant because recent work has suggested NAc-related projections are significant in treating patients with depression using DBS[10,82]. It will be interesting to investigate possible anatomical and causal linkages between the mPFC–DRN[76] and VTA–NAc[75] pathways.

The vegetative symptoms of psychiatric disease are those linked to basic physiological functions, including appetite, sleep–wake balance and sexual behaviour. These symptoms are individually variable during depressive episodes, but in aggregate are common, debilitating and constitute core criteria for diagnosis of major depression. The initial optogenetic control of behaviour in mammals was in fact conducted on sleep–wake behaviour through the hypocretin neurons of the lateral hypothalamus; certain patterns of activity delivered to these cells in mice were found to favour the transition from sleep to wake[40]. It was later found that driving noradrenaline neurons in the locus coeruleus (LC) exerted a rapid and profound effect on awakening[83,84], and by driving hypocretin neurons with an excitatory opsin while simultaneously inhibiting LC noradrenaline neurons with an inhibitory opsin, it was observed that the effect of hypocretin neurons on awakening[40,84,85] could be accounted for by projections to the LC. Early applications of optogenetics to another vegetative category, feeding behaviour, have included defining (and modulating) the reward value of feeding (an important step towards understanding how the reward value of feeding could become altered in the setting of psychiatric disease). Precisely tuned control over the value of nutrient consumption was achieved using optogenetic drive of DA neurons in the VTA, which shifted the value of one class of nutrient to be greater than the highly appetitive value of sucrose[86]. Later papers used optogenetic projection targeting to evoke strikingly excessive feeding behaviour with relevance to depression, eating disorders and obesity by driving agouti-related peptide (AgRP) neurons that project from the arcuate nucleus to the paraventricular hypothalamus (PVH)[87] or by driving the GABAergic projection from the BNST to the lateral hypothalamus[88].

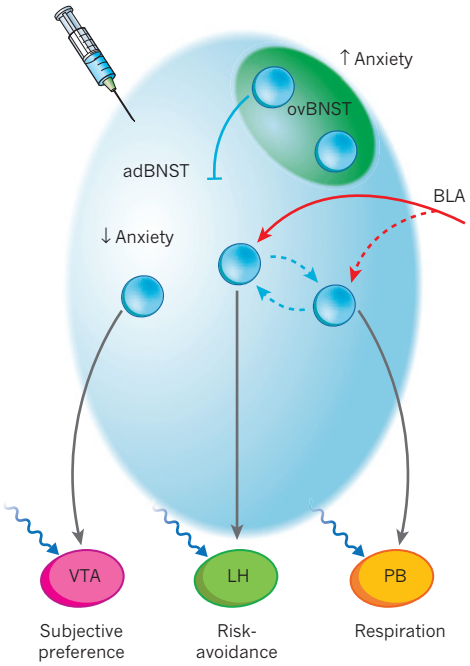


Figure 2 | Controlling projections in the intact mammalian brain. Behavioural state assembly and complexity can be assessed with projection targeting. Functional organization of the bed nucleus of the stria terminalis (BNST) circuitry is shown. The oval nucleus of the BNST (ovBNST) is anxiogenic, and the anterodorsal BNST (adBNST) is anxiolytic[68], as is the projection from the BLA (red arrows); the ovBNST may act to increase anxiety by inhibiting the adBNST or by independently influencing downstream structures. The adBNST projects to the ventral tegmental area (VTA), lateral hypothalamus (LH) and parabrachial nucleus (PB); each of these projections decreases a distinct feature of anxiety expression, and coordinated recruitment of these subpopulations may be implemented by recurrent circuitry in adBNST[68]. Projections are assessed by introduction of opsin (needle) into the adBNST, and independent targeting of fibre-optic interfaces depicted for each candidate downstream target (blue arrows). Here (as with medial prefrontal cortex projections and VTA projections) little mixed-site collateralization from single neurons is seen[68]. Adapted with permission from ref. 68.

Hedonic and anhedonic behaviours, beyond their relevance to feeding alone, are of substantial significance in psychiatric disease; anhedonia is a formal diagnostic criterion for depression, distinct from vegetative and motivational criteria. At the other end of the spectrum, excessive hedonic behaviour manifests in conditions relating to substance abuse and dependency, as well as in manic and hypomanic conditions. Dysfunction in responding to rewarding or appetitive stimuli can be studied in animals by determining to what extent behaviour can be positively conditioned by, or to what extent animals will work for, the stimuli in disease-like states[74]. Over the past five years, applications of optogenetic projection targeting under physiological conditions and in disease models have begun to illuminate the causal structural dynamics of hedonia and anhedonia.

An initial step came with the optogenetic drive of VTA DA neurons, to causally test whether different patterns of spikes in these defined cells could be appetitive for mammals[89]. In freely moving mice, high-frequency 'phasic' bursts of spikes delivered to VTA DA neurons were found to potently drive place-preference conditioning, and, interestingly, seemed preferable to mice compared with the same number of spikes delivered at a lower 'tonic' frequency[89] (highlighting the importance for optogenetics of eliciting precise temporal dynamics in the targeted cells). Although these initial experiments were carried out in *Th*-Cre transgenic mice to allow the targeting of VTA DA neurons with Cre-dependent opsin-expressing adeno-associated virus (AAV) vectors, in later work[81] rats were used as experimental subjects for robust assessment of the extent to which animals will work for a stimulus or reward. *Th*-Cre transgenic rats designed and created for this type of experiment were observed to execute many thousands of nose pokes per day (a test for experimental subject-initiated, motivated behaviour) to obtain optogenetically delivered phasic spike bursts in VTA DA neurons; furthermore, projection targeting revealed that rats will work for activity in VTA-DA cells that are even more precisely defined, namely the subset of these neurons projecting to the NAc[81]. It is worth noting that clinical DBS findings have pointed to the NAc in anhedonic symptoms of major depression[90,91].

Optogenetic projection targeting has identified a number of additional afferents to the NAc that also modulate reward-related behaviour or appetitive conditioning, in some cases connected to psychiatric-disease-related symptoms. Not only will animals work to receive stimulation of the defined BLA projection to the NAc, but also optogenetic inhibition of this projection creates a depressive-like phenotype of reduced motivation to obtain a normally appetitive sucrose solution[92]. Subsequent work has implicated excitatory afferents to the NAc from the ventral hippocampus (vH) and mPFC in promoting appetitive conditioning and reward-related behaviour[93]; dopamine type-1 (D1) receptors in NAc were implicated, and the vH pathway to the NAc was found to be potentiated by cocaine intake[93], consistent with the earlier finding that D1 receptors in the NAc are important for appetitive cocaine responses[94]. Additional cocaine-related projection targeting studies have illuminated the relevance of PFC–NAc pathways; inhibition of the projection from the PFC to the NAc blocked cocaine-seeking behaviour in rats[95], and specific inhibition of mPFC inputs to the NAc by optogenetic depotentiation elicited a reduction in cocaine-induced locomotor sensitization[96]. Together, these studies have elucidated fundamental aspects of causal structural dynamics that underlie abused-substance-related hedonic behaviours.

Recently, much important information has also come to light concerning the causal role of the projections into the VTA itself, which is relevant to hedonic and anhedonic behavioural states. An opsin-carrying rabies virus was delivered into the VTA; projection targeting was then achieved by delivering light to distinct upstream cell-body regions[97]. Conditioned place preference was seen when the projection from the laterodorsal tegmentum (LDT) to the VTA was driven (interestingly these projections synapse preferentially onto NAc-projecting outgoing VTA DA neurons)[97]. By contrast, conditioned aversion was observed when the projection to the VTA was driven from the LHb[97] — the same structure that supported a depressive-like passive-coping response in the FST when mPFC-to-LHb projections were driven[76]. In a complementary (and concordant) projection targeting set of results[98], LHb projections to the VTA were optically driven by transducing the LHb with a channelrhodopsin and illuminating the VTA; this intervention was also aversive (and in fact antagonized positive reinforcement)[98]. Interestingly, the aversive LHb projections to the VTA seem to favour, as synaptic targets, DA neurons that project, not to the NAc, but instead to the mPFC[97], perhaps outlining a passive–aversive loop of influence definable by optogenetic projection targeting, spanning the brain and schematized as mPFC–LHb–VTA–mPFC. There may be antagonism or competition for influence between appetitive and aversive projection networks across the brain, as the aversive LHb also seems to drive GABAergic cells in the rostromedial tegmental nucleus (RMTg) that in turn inhibit putatively behaviourally activating or appetitive VTA neurons projecting to the NAc (lateral shell region)[97].

Other influences beyond that of mPFC converge onto the LHb; for example, in rats, projection targeting was used to drive the glutamatergic entopeduncular nucleus to LHb projection[99], resulting in conditioned place aversion. The potential for complex regulation is substantial and, given the spatial overlap of the passive or aversive circuitry with behaviourally activating or appetitive influences at nearly every step (notably in the mPFC and VTA), no doubt both the anatomical precision of tract targeting and the type of behavioural challenge delivered will determine which competing pathway experimentally predominates. As an example of anatomical complexity, projections to the VTA can either promote or inhibit reward, depending on location and cell type of fibre origin in the anterodorsal or ventral BNST, respectively[68,100]. And as an example of behavioural history complexity, severe and acute social-defeat stressors seem to favour recruitment of an aversion-related DA cell population in VTA[101], in contrast to chronic mild stressors that favour recruitment of a behaviourally activating or appetitive DA VTA population[75]. Such complexity is expected in the mammalian brain; these experiments have provided only the first insights into the causal impact of defined spike patterns in specified cells and projections on mammalian hedonic behaviour.

Outlook

What is the significance of the observation that projection-specific dynamics operate naturally as control levers that are relevant to psychiatry? Careful interpretation is important here because a tested projection presumably does not subserve only the role implicated (Table 1), nor need all, or even most, of the fibres within the projecting tract contribute to the behavioural effects seen; of course, these considerations apply to any experimental intervention. The sign and direction of the net behavioural change captures only one resultant of the causal influence of the projection tested. Still, as already noted, in many cases a projection-specific dynamical modulation delivers more efficacious elicitation or correction of psychiatry-related adaptive or maladaptive behaviours compared with region-specific dynamical modulation (despite probably recruiting fewer targeted circuit elements). This pattern even extends beyond the psychiatric-disease-related domain; for example, recently, more precisely tuned behavioural changes were obtained through corticostriatal optogenetic projection targeting compared with direct stimulation in a challenging auditory behaviour task[48].

One interpretation is that greater experimental efficacy may be obtained through the directional and intersectional specificity (Box 1) inherent to AAV opsin transduction integrated with spatially separated fibre-optic-based projection targeting, which minimizes the generation of conflicting signals (such as push–pull on the same behaviour, as was observed in the mPFC FST). In addition, projection targeting allows direct control to only be exerted over cells that have long-range projections, without directly perturbing purely locally connected neurons that may be engaged in active circuit computations; although projecting cells can also be involved in computation and representation of information, these roles may be tied closely enough to the projection wiring itself that

delivery or inhibition of experimenter-provided spikes can be meaningful for modulating complex behaviours. But beyond such experimental considerations, the mammalian brain seems to be set up to allow natural behavioural state features to be tuned by projection dynamics (Fig. 2 and Table 1), and many psychiatric symptoms (such as the aversive or negative quality of the anxious state, which is dysregulated to the point of morbidity in anxiety disorders) may be extreme manifestations of these natural features. As already described, multiple-feature natural states (for anxiety, these can include respiratory and behavioural changes) can be managed from a centralized node (for example, the BNST) that assembles these components through outgoing projections[68] (Fig. 2). These projections may set the gain for execution of each feature, so that turning up or down the activity of these defined projections accesses a natural anatomically defined circuit signal for feature (or symptom) intensity.

This causal structural dynamics perspective contributes to our understanding of natural behaviour, but may also have clinical insight implications. Individual features of a behavioural state can be debilitating but not necessarily the overall state itself. For example, it may be appropriate to elevate both the respiratory rate and alertness to danger in a new environment, but only if the additional feature of subjective aversion is strongly recruited does a pathological state resembling clinical anxiety result. The clinical problem can be a feature of the state, not the state itself, and therefore identification of projection-specific dynamics that are relevant to disease has implications for our understanding of aversive, appetitive or maladaptive features that are more closely linked to clinically relevant morbidity than the overall behavioural state itself. Improved brain stimulation treatments[6,102] may eventually result from guiding focal interventions to brain locations where the identified symptom-related projection is most resolved from other projections; here integration of causal information derived from optogenetics may dovetail with patient-specific anatomical information from diffusion magnetic resonance imaging (dMRI)-based tract mapping and modelling[103]. Investigation of the synaptic plasticity rules governing optogenetically defined projections may also help us to elicit stable effects from brief interventions (which could be delivered through functional magnetic resonance imaging (fMRI)-guided trans-cranial magnetic stimulation (TMS)[104,105]). Molecular characterization of identified projections with techniques[61] for high-content identification of expressed transcripts and proteins may facilitate the development of more specific projection-informed and targeted drug therapies. And it will be interesting to explore whether separate efficacious white-matter DBS target locations for a given disorder[106] actually share a common projection component that might be identified by dMRI or CLARITY in human or animal settings[107–109]. Moreover, animal-model identification of causally important projections may guide seed placement for clinical dMRI mapping, even as dMRI maps linked to clinical symptoms in turn feed back to inform experimental target selection in animals and guide dynamical modelling.

In principle, modulation of connectivity properties can exert a highly sensitive and consequential influence (in certain models) over network dynamics and information flow[110–112], and simulated lesioning of locations (nodes or vertices in brain graphs) wired with different statistics (projection or edge properties) can give rise to different effects on network dynamics and connectivity[113]. However, as is well-recognized by graph theorists in neuroscience[114,115], it is not completely clear what biological scale these edges and vertices might most usefully represent, and in the course of generating hypotheses and valuable insights into network complexity and connectivity, current graph-theoretic approaches have considered spatial scales that range from cells to macroscopic subdivisions of the human brain[116]. Linking this field with causal projection-dynamics information arising from optogenetics may help to concretize the nature and scale of graph components and network measures, such as those describing node connectivity[113] (for example betweenness centrality). The physical understanding provided by projection-dynamic studies could also help us to interpret the effects of other circuit phenomena on psychiatric symptoms. For example, local E/I balance shifted by neurochemical influences or medications could alter recurrent feedback within a brain region such as the mPFC or extended amygdala (Box 1), and, in doing so, determine the extent of coordinated recruitment of behavioural state features or symptoms governed by specific outgoing projections[68].

Although broad avenues and opportunities exist for leveraging insights from causal structural dynamics findings, there are also limitations. Certainly, the challenge of projection complexity should be appreciated. Isolated tracts can control multiple behaviours, individual axons within a tract could differentially branch and affect multiple downstream regions in a symptom-relevant fashion, and individual axons could also deliver differentially complex and non-canonical combinations of neurotransmitters[117] that are potentially relevant to the diversity of symptoms seen in single discrete psychiatric diseases. From the perspective of psychiatric disease, it is also important to underscore that knowledge of causal projection-dynamics phenomena will be largely limited to those accessible with optogenetics and therefore will tend to exclude symptom domains that lack widely accepted animal models[74], extending to the delusions, hallucinations and disordered thinking of schizophrenia. These less-accessible symptom domains still probably involve dysfunctional projections or wiring[110,118–120] and are clinically known to cluster with other behavioural phenomena that are accessible in animal models (such as social withdrawal and stereotypies), which may therefore share related underlying circuit-dynamics manifestations; improved multidimensional and high-throughput behavioural quantification in the setting of simultaneous physiological readouts will facilitate detection of such unifying themes.

But even reliable and standard animal behavioural measures are subject to controversy and differences in interpretation. One opportunity in causal structural dynamics research going forward may therefore be to improve disease-model validation and to refine standard animal behavioural measures, partly through integration of projection targeting behavioural research with clinical projection research (such as recent large-scale human connectome efforts using dMRI, involving tractography data linked to symptoms). But in the end, no animal model will fully capture the ontogeny, pathophysiology and symptomatic complexity of human psychiatric disease. Therefore, it is useful to maintain attention not on diseases, but on restricted symptom domains. This approach is ideal for animal modelling, well-suited to the specificity of optogenetic projection targeting and well-aligned with the symptom-treatment approach of real-world clinical psychiatry.

Psychiatric symptom domains, involving dynamical and distributed performance changes without frank regional cell loss, may in this regard share certain principles with developmental and evolutionary changes in neural circuitry. For example, over time certain environments and situations may come to assume a different significance for an organism, and it will be important to adaptably recruit different features of physiology to create a behavioural state more optimized for this shifting significance — without requiring a fundamental redesign of the neural coding of behaviour or representation of the environment itself. This perspective suggests why certain classes of projections may be particularly plastic — or designed for variability — in a way that does not affect the neural code per se that corresponds to a motor plan or representation of sensory data, but rather the magnitude and valence of a behavioural state feature. This flexibility in projection functionality may in turn help to explain the prevalence of psychiatry-related structural dynamics in the mammalian population, representing hotspots for adaptive or maladaptive alteration in symptoms analogous to chromosomal hotspots for mutation relevant to cancer. Among other useful properties, projection dynamics may not only represent concrete physical manifestations of psychiatric symptoms, but may also help to contribute to modularity and flexibility of behavioural states during development and evolution. ■

Received 31 July; accepted 8 November 2013.

1. Akil, H. *et al*. The future of psychiatric research: genomes and neural circuits. *Science* **327,** 1580–1581 (2010).
2. Deisseroth, K. Optogenetics and psychiatry: applications, challenges, and

opportunities. *Biol. Psychiatry* **71,** 1030–1032 (2012).
3. Abbott, A. Novartis to shut brain research facility. *Nature* **480,** 161–162 (2011).
4. Deisseroth, K. Controlling the brain with light. *Sci. Am.* **303,** 48–55 (2010).
5. Maks, C. B., Butson, C. R., Walter, B. L., Vitek, J. L. & McIntyre, C. C. Deep brain stimulation activation volumes and their association with neurophysiological mapping and therapeutic outcomes. *J. Neurol. Neurosurg. Psychiatry* **80,** 659–666 (2009).
6. Holtzheimer, P. E. & Mayberg, H. S. Deep brain stimulation for psychiatric disorders. *Annu. Rev. Neurosci.* **34,** 289–307 (2011).
This is a recent comprehensive review of DBS targets and effects in psychiatry.
7. Oluigbo, C. O., Salma, A. & Rezai, A. R. Deep brain stimulation for neurological disorders. *IEEE Rev. Biomed. Eng.* **5,** 88–99 (2012).
8. Benabid, A. L. & Torres, N. New targets for DBS. *Parkinsonism Relat. Disord*. **18** (Suppl 1), 21–23 (2012).
9. Goodman, W. K. & Alterman, R. L. Deep brain stimulation for intractable psychiatric disorders. *Annu. Rev. Med.* **63,** 511–524 (2012).
10. Bewernick, B. H., Kayser, S., Sturm, V. & Schlaepfer, T. E. Long-term effects of nucleus accumbens deep brain stimulation in treatment-resistant depression: evidence for sustained efficacy. *Neuropsychopharmacology* **37,** 1975–1985 (2012).
11. Chan, D. T. *et al*. Complications of deep brain stimulation: a collective review. *Asian J. Surg.* **32,** 258–263 (2009).
12. Birdno, M. J. & Grill, W. M. Mechanisms of deep brain stimulation in movement disorders as revealed by changes in stimulus frequency. *Neurotherapeutics* **5,** 14–25 (2008).
13. Gradinaru, V., Mogri, M., Thompson, K. R., Henderson, J. M. & Deisseroth, K. Optical deconstruction of parkinsonian neural circuitry. *Science* **324,** 354–359 (2009).
This initial study of optogenetic projection control in modulating behaviour targeted afferents to the subthalamic nucleus that influence parkinsonian symptoms.
14. Andrews-Zwilling, Y. *et al*. Hilar GABAergic interneuron activity controls spatial learning and memory retrieval. *PLoS ONE* **7,** e40555 (2012).
15. Rubenstein, J. L. & Merzenich, M. M. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav.* **2,** 255–267 (2003).
16. Yizhar, O. *et al*. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477,** 171–178 (2011).
17. Chao, H. T. *et al*. Dysfunction in GABA signalling mediates autism-like stereotypies and Rett syndrome phenotypes. *Nature* **468,** 263–269 (2010).
18. Akam, T., Oren, I., Mantoan, L., Ferenczi, E. & Kullmann, D. M. Oscillatory dynamics in the hippocampus support dentate gyrus–CA3 coupling. *Nature Neurosci.* **15,** 763–768 (2012).
19. Blumhagen, F. *et al*. Neuronal filtering of multiplexed odour representations. *Nature* **479,** 493–498 (2011).
20. Cardin, J. A. *et al*. Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459,** 663–667 (2009).
21. Carlén, M. *et al*. A critical role for NMDA receptors in parvalbumin interneurons for gamma rhythm induction and behavior. *Mol. Psychiatry* **17,** 537–548 (2012).
22. Figee, M. *et al*. Deep brain stimulation restores frontostriatal network activity in obsessive-compulsive disorder. *Nature Neurosci.* **16,** 386–387 (2013).
23. Rho, Y. A., McIntosh, R. A. & Jirsa, V. K. Synchrony of two brain regions predicts the blood oxygen level dependent activity of a third. *Brain Connect.* **1,** 73–80 (2011).
24. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* **459,** 698–702 (2009).
25. Tiesinga, P. H. & Sejnowski, T. J. Mechanisms for phase shifting in cortical networks and their role in communication through coherence. *Front. Hum. Neurosci.* **4,** 196 (2010).
26. Jirsa, V. K. Connectivity and dynamics of neural information processing. *Neuroinformatics* **2,** 183–204 (2004).
27. Stroh, A. *et al*. Making waves: initiation and propagation of corticothalamic Ca^{2+} waves *in vivo*. *Neuron* **77,** 1136–1150 (2013).
28. Airan, R. D. *et al*. High-speed imaging reveals neurophysiological links to behavior in an animal model of depression. *Science* **317,** 819–823 (2007).
29. Gutierrez, G. J., O'Leary, T. & Marder, E. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron* **77,** 845–858 (2013).
30. Deisseroth, K. Optogenetics. *Nature Methods* **8,** 26–29 (2011).
31. Packer, A. M., Roska, B. & Hausser, M. Targeting neurons and photons for optogenetics. *Nature Neurosci.* **16,** 805–815 (2013).
32. Fenno, L., Yizhar, O. & Deisseroth, K. The development and application of optogenetics. *Annu. Rev. Neurosci.* **34,** 389–412 (2011).
33. Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M. & Deisseroth, K. Optogenetics in neural systems. *Neuron* **71,** 9–34 (2011).
34. Gradinaru, V. *et al*. Targeting and readout strategies for fast optical neural control *in vitro* and *in vivo*. *J. Neurosci.* **27,** 14231–14238 (2007).
35. Li, Q. *et al*. Therapeutic deep brain stimulation in Parkinsonian rats directly influences motor cortex. *Neuron* **76,** 1030–1041 (2012).
36. Tye, K. M. *et al*. Amygdala circuitry mediating reversible and bidirectional control of anxiety. *Nature* **471,** 358–362 (2011).
This was the initial study targeting specific projections in behaviour; gain- or loss-of-function was delivered by optogenetic excitation or inhibition to a specific amygdala projection, with the resulting bidirectional expression of anxiety-related behaviours.
37. Akerboom, J. *et al*. Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. *Front. Mol. Neurosci.* **6,** 2 (2013).
38. Kuhn, B., Ozden, I., Lampi, Y., Hasan, M. T. & Wang, S. S. An amplified promoter system for targeted expression of calcium indicator proteins in the cerebellar cortex. *Front. Neural Circuits* **6,** 49 (2012).
39. Saunders, A., Johnson, C. A. & Sabatini, B. L. Novel recombinant adeno-associated viruses for Cre activated and inactivated transgene expression in neurons. *Front. Neural Circuits* **6,** 47 (2012).
40. Adamantidis, A. R., Zhang, F., Aravanis, A. M., Deisseroth, K. & de Lecea, L. Neural substrates of awakening probed with optogenetic control of hypocretin neurons. *Nature* **450,** 420–424 (2007).
41. Zhang, F. *et al*. Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. *Nature Protocols* **5,** 439–456 (2010).
42. Osakada, F. *et al*. New rabies virus variants for monitoring and manipulating activity and gene expression in defined neural circuits. *Neuron* **71,** 617–631 (2011).
43. Antinone, S. E. & Smith, G. A. Retrograde axon transport of herpes simplex virus and pseudorabies virus: a live-cell comparative analysis. *J. Virol.* **84,** 1504–1512 (2010).
44. Miyamichi, K. *et al*. Cortical representations of olfactory input by trans-synaptic tracing. *Nature* **472,** 191–196 (2011).
45. Wickersham, I. R. et al. Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53,** 639–647 (2007).
References 44 and 45 describe rabies-based tools used to define and study afferent projections to cell populations *in vivo*.
46. Lima, S. Q., Hromadka, T., Znamenskiy, P. & Zador, A. M. PINP: a new method of tagging neuronal populations for identification during *in vivo* electrophysiological recording. *PLoS ONE* **4,** e6099 (2009).
47. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482,** 85–88 (2012).
48. Znamenskiy, P. & Zador, A. M. Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* **497,** 482–485 (2013).
49. Lichtman, J. W. & Denk, W. The big and the small: challenges of imaging the brain's circuits. *Science* **334,** 618–623 (2011).
50. Osten, P. & Margrie, T. W. Mapping brain circuitry with a light microscope. *Nature Methods* **10,** 515–523 (2013).
51. Bock, D.D. *et al*. Network anatomy and *in vivo* physiology of visual cortical neurons. *Nature* **471,** 177–182 (2011).
52. Briggman, K.L., Helmstaedter, M. & Denk, W. Wiring specificity in the direction-selectivity circuit of the retina. *Nature* **471,** 183–188 (2011).
53. Micheva, K. D. & Smith, S. J. Array tomography: a new tool for imaging the molecular architecture and ultrastructure of neural circuits. *Neuron* **55,** 25–36 (2007).
54. Hama, H. *et al*. Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain. *Nature Neurosci.* **14,** 1481–1488 (2011).
55. Dodt, H. U. *et al*. Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. *Nature Methods* **4,** 331–336 (2007).
56. Ertürk, A. *et al*. Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* **7,** 1983–1995 (2012).
57. Ke, M.-T., Fujimoto, S. & Imai, T. SeeDB: a simple and morphology-preserving optical clearing agent for neuronal circuit reconstruction. *Nature Neurosci.* **16,** 1154–1161 (2013).
58. Kuwajima, T. *et al*. ClearT: a detergent- and solvent-free clearing method for neuronal and non-neuronal tissue. *Development* **140,** 1364–1368 (2013).
59. Bohland, J. W. *et al*. A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. *PLoS Comput. Biol*. **5,** e1000334 (2009).
60. Chung, K. *et al*. Structural and molecular interrogation of intact biological systems. *Nature* **497,** 332–337 (2013).
This article reports a chemical engineering method for visualizing and labelling projections in the intact brain.
61. Chung, K. & Deisseroth, K. CLARITY for mapping the nervous system. *Nature Methods* **10,** 508–513 (2013).
62. Regehr, W. G. & Tank, D. W. Selective fura-2 loading of presynaptic terminals and nerve cell processes by local perfusion in mammalian brain slice. *J. Neurosci. Methods* **37,** 111–119 (1991).
63. Tye, K. M. & Deisseroth, K. Optogenetic investigation of neural circuits underlying brain disease in animal models. *Nature Rev. Neurosci.* **13,** 251–266 (2012).
64. Ciocchi, S. *et al*. Encoding of conditioned fear in central amygdala inhibitory circuits. *Nature* **468,** 277–282 (2010).
65. Haubensak, W. *et al*. Genetic dissection of an amygdala microcircuit that gates conditioned fear. *Nature* **468,** 270–276 (2010).
66. Iwata, J. & LeDoux, J. E. Dissociation of associative and nonassociative concomitants of classical fear conditioning in the freely behaving rat. *Behav. Neurosci.* **102,** 66–76 (1988).
67. Johansen, J. P. *et al*. Optical activation of lateral amygdala pyramidal cells instructs associative fear learning. *Proc. Natl Acad. Sci. USA* **107,** 12692–12697 (2010).
68. Kim, S. Y. *et al*. Diverging neural pathways assemble a behavioural state from separable features in anxiety. *Nature* **496,** 219–223 (2013).
This study reports the optogenetic decomposition of a behavioural state into component features by projection-targeting-based recruitment of separable anxiety-related features.

69. Fineberg, N. A. *et al*. Probing compulsive and impulsive behaviours, from animal models to endophenotypes: a narrative review. *Neuropsychopharmacology* **35,** 591–604 (2010).
70. Ahmari, S. E. *et al*. Repeated cortico-striatal stimulation generates persistent OCD-like behavior. *Science* **340,** 1234–1239 (2013).
71. Burguière, E., Monteiro, P., Feng, G. & Graybiel, A. M. Optogenetic stimulation of lateral orbitofronto-striatal pathway suppresses compulsive behaviors. *Science* **340,** 1243–1246 (2013).
72. Stopper, C. M., Green, E. B. & Floresco, S. B. Selective involvement by the medial orbitofrontal cortex in biasing risky, but not impulsive, choice. *Cereb. Cortex* **24,** 154–162 (2014).
73. Krishnan, V. & Nestler, E. J. Animal models of depression: molecular perspectives. *Curr. Topics Behav. Neurosci.* **7,** 121–147 (2011).
74. Nestler, E. J. & Hyman, S. E. Animal models of neuropsychiatric disorders. *Nature Neurosci.* **13,** 1161–1169 (2010).
75. Tye, K. M. *et al*. Dopamine neurons modulate neural encoding and expression of depression-related behaviour. *Nature* **493,** 537–541 (2013).
76. Warden, M. R. *et al*. A prefrontal cortex-brainstem neuronal projection that controls response to behavioural challenge. *Nature* **492,** 428–432 (2012).
This article describes optogenetic projection-targeting-based recruitment of prefrontal pathways favouring active-coping or passive-coping behavioural patterns relevant to depression.
77. Voon, V. *et al*. Dopamine agonists and risk: impulse control disorders in Parkinson's disease. *Brain* **134,** 1438–1446 (2011).
78. Young, J. W., van Enkhuizen, J., Winstanley, C. A. & Geyer, M. A. Increased risk-taking behavior in dopamine transporter knockdown mice: further support for a mouse model of mania. *J. Psychopharmacol.* **25,** 934–943 (2011).
79. Farrell, S. M., Tunbridge, E. M., Braeutigam, S. & Harrison, P. J. COMT Val^{158} Met genotype determines the direction of cognitive effects produced by catechol-*O*-methyltransferase inhibition. *Biol. Psychiatry* **71,** 538–544 (2012).
80. Foti, D. & Hajcak, G. Genetic variation in dopamine moderates neural response during reward anticipation and delivery: evidence from event-related potentials. *Psychophysiology* **49,** 617–626 (2012).
81. Witten, I. B. *et al*. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron* **72,** 721–733 (2011).
82. Bewernick, B. H. *et al*. Nucleus accumbens deep brain stimulation decreases ratings of depression and anxiety in treatment-resistant depression. *Biol. Psychiatry* **67,** 110–116 (2010).
83. Carter, M. E. & de Lecea, L. Optogenetic investigation of neural circuits *in vivo*. *Trends Mol. Med.* **17,** 197–206 (2011).
84. Carter, M. E. *et al*. Tuning arousal with optogenetic modulation of locus coeruleus neurons. *Nature Neurosci.* **13,** 1526–1533 (2010).
85. Carter, M. E. *et al*. Mechanism for hypocretin-mediated sleep-to-wake transitions. *Proc. Natl Acad. Sci. USA* **109,** E2635–E2644 (2012).
86. Domingos, A. I. *et al*. Leptin regulates the reward value of nutrient. *Nature Neurosci.* **14,** 1562–1568 (2011).
87. Atasoy, D., Betley, J. N., Su, H. H. & Sternson, S. M. Deconstruction of a neural circuit for hunger. *Nature* **488,** 172–177 (2012).
88. Jennings, J. H., Rizzi, G., Stamatakis, A. M., Ung, R. L. & Stuber, G. D. The inhibitory circuit architecture of the lateral hypothalamus orchestrates feeding. *Science* **341,** 1517–1521 (2013).
89. Tsai, H. C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324,** 1080–1084 (2009).
90. Schlaepfer, T. E. *et al.* Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. *Neuropsychopharmacology* **33,** 368–377 (2008).
91. Grubert, C. *et al*. Neuropsychological safety of nucleus accumbens deep brain stimulation for major depression: effects of 12-month stimulation. *World J. Biol. Psychiatry* **12,** 516–527 (2011).
92. Stuber, G. D. *et al.* Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* **475,** 377–380 (2011).
Early projection-targeting study in mammalian appetitive and aversive conditioning, demonstrating bidirectional control of the BLA to nucleus accumbens projection using ChR2 and eNpHR3.0.
93. Britt, J. P. *et al*. Synaptic and behavioral profile of multiple glutamatergic inputs to the nucleus accumbens. *Neuron* **76,** 790–803 (2012).
94. Lobo, M. K. *et al*. Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science* **330,** 385–390 (2010).
95. Stefanik, M. T. *et al*. Optogenetic inhibition of cocaine seeking in rats. *Addict. Biol.* **18,** 50–53 (2013).
96. Pascoli, V., Turiault, M. & Luscher, C. Reversal of cocaine-evoked synaptic potentiation resets drug-induced adaptive behaviour. *Nature* **481,** 71–75 (2012).
97. Lammel, S. *et al*. Input-specific control of reward and aversion in the ventral tegmental area. *Nature* **491,** 212–217 (2012).
This paper describes the behavioural control of appetitive and aversive conditioning by targeting distinct incoming projections (from LDT or LHb) into the VTA.
98. Stamatakis, A. M. & Stuber, G. D. Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance. *Nature Neurosci.* **15,** 1105–1107 (2012).
99. Shabel, S. J., Proulx, C. D., Trias, A., Murphy, R. T. & Malinow, R. Input to the lateral habenula from the basal ganglia is excitatory, aversive, and suppressed by serotonin. *Neuron* **74,** 475–481 (2012).
100. Jennings, J. H. *et al*. Distinct extended amygdala circuits for divergent motivational states. *Nature* **496,** 224–228 (2013).
101. Chaudhury, D. *et al*. Rapid regulation of depression-related behaviours by control of midbrain dopamine neurons. *Nature* **493,** 532–536 (2013).
102. Lozano, A. M. & Lipsman, N. Probing and regulating dysfunctional circuits using deep brain stimulation. *Neuron* **77,** 406–424 (2013).
103. Deco, G., Senden, M. & Jirsa, V. How anatomy shapes dynamics: a semi-analytical study of the brain at rest by a simple spin model. *Front. Comput. Neurosci.* **6,** 68 (2012).
104. Li, X. *et al*. Using interleaved transcranial magnetic stimulation/functional magnetic resonance imaging (fMRI) and dynamic causal modeling to understand the discrete circuit specific changes of medications: lamotrigine and valproic acid changes in motor or prefrontal effective connectivity. *Psychiatry Res.* **194,** 141–148 (2011).
105. Chen, A. C. Causal interactions between front-parietal central executive and default-mode networks in humans. *Proc. Natl Acad. Sci. USA* **110,** 19944–19949 (2013).
106. Gutman, D. A., Holtzheimer, P. E., Behrens, T. E., Johansen-Berg, H. & Mayberg, H. S. A tractography analysis of two deep brain stimulation white matter targets for depression. *Biol. Psychiatry* **65,** 276–282 (2009).
107. Toga, A. W., Ambach, K., Quinn, B., Hutchin, M. & Burton, J. S. Postmortem anatomy from cryosectioned whole human brain. *J. Neurosci. Methods* **54,** 239–252 (1994).
108. Rauschning, W. Surface cryoplaning. A technique for clinical anatomical correlations. *Ups. J. Med. Sci.* **91,** 251–255 (1986).
109. Amunts, K. *et al*. BigBrain: an ultrahigh-resolution 3D human brain model. *Science* **340,** 1472–1475 (2013).
110. Cabral, J., Kringelbach, M. L. & Deco, G. Functional graph alterations in schizophrenia: a result from a global anatomic decoupling? *Pharmacopsychiatry* **45** (Suppl 1), 57–64 (2012).
111. Pinotsis, D. A., Hansen, E., Friston, K. J. & Jirsa, V. K. Anatomical connectivity and the resting state activity of large cortical networks. *Neuroimage* **65,** 127–138 (2013).
112. Sporns, O. The non-random brain: efficiency, economy, and complex dynamics. *Front. Comput. Neurosci.* **5,** 5 (2011).
113. Jirsa, V. K., Sporns, O., Breakspear, M., Deco, G. & McIntosh, A. R. Towards the virtual brain: network modeling of the intact and the damaged brain. *Arch. Ital. Biol.* **148,** 189–205 (2010).
114. Sporns, O. The human connectome: a complex network. *Ann. NY Acad. Sci.* **1224,** 109–125 (2011).
115. Bullmore, E. T. & Bassett, D. S. Brain graphs: graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* **7,** 113–140 (2011).
116. Fornito, A. & Bullmore, E. T. Connectomic intermediate phenotypes for psychiatric disorders. *Front. Psychiatry* **3,** 32 (2012).
117. Tritsch, N. X., Ding, J. B. & Sabatini, B. L. Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. *Nature* **490,** 262–266 (2012).
118. Fitzsimmons, J., Kubicki, M. & Shenton, M. E. Review of functional and anatomical brain connectivity findings in schizophrenia. *Curr. Opin. Psychiatry* **26,** 172–187 (2013).
119. Lynall, M. E. *et al*. Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* **30,** 9477–9487 (2010).
120. Maher, B. J. & LoTurco, J. J. Disrupted-in-schizophrenia (DISC1) functions presynaptically at glutamatergic synapses. *PLoS ONE* **7,** e34053 (2012).
121. Felix-Ortiz, A. C. *et al*. BLA to vHPC inputs modulate anxiety-related behaviors. *Neuron* **79,** 658–664 (2013).
122. Gunaydin, L. *et al*. Real-time optical measurement of projection activity: dynamics of genetically- and anatomically-defined neuronal afferents predict social behavior in freely moving mice. Program No. 872.04. 2013 *Neuroscience Meeting Planner*. (Society for Neuroscience, 2013).
123. Stamatakis, A. M. A unique population of ventral tegmental area neurons inhibits the lateral habenula to promote reward. *Neuron* **20,** 1039–1053 (2013).
124. Gradinaru, V. *et al*. Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141,** 154–165 (2010).
125. Petreanu, L., Huber, D., Sobczyk, A. & Svoboda, K. Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. *Nature Neurosci.* **10,** 663–668 (2007).
126. Mattis, J. *et al*. Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nature Methods* **9,** 159–172 (2012).
127. Zalocusky, K. & Deisseroth, K. Optogenetics in the behaving rat: integration of diverse new technologies in a vital animal model. *Optogenetics* **2013,** 1–17 (2013).

Acknowledgements I am deeply indebted to my patients over the years for their insight, perseverance and strength in working to convey the most complex and nearly inarticulatable inner thought processes and feelings associated with severe psychiatric disease. I am also grateful to my entire laboratory for support, as well as to V. Sohal, A. Schatzberg, H. Mayberg, R. Malenka, A. Etkin, L. Grosenick, A. Kreitzer, T. Insel, M. Warden, M. Zelikowsky and E. Ferenczi for comments and discussions over the years. K.D. has been supported by the Wiegers Family Fund, NARSAD, NIMH, NIDA, DARPA, the Keck Foundation, the McKnight Foundation, the Yu, Snyder and Woo Foundations, and the Gatsby Charitable Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/gmrxho. Correspondence should be addressed to K.D. (deissero@stanford.edu).

REVIEW

doi:10.1038/nature12983

Interneuron cell types are fit to function

Adam Kepecs[1] & Gordon Fishell[2]

Understanding brain circuits begins with an appreciation of their component parts — the cells. Although GABAergic interneurons are a minority population within the brain, they are crucial for the control of inhibition. Determining the diversity of these interneurons has been a central goal of neurobiologists, but this amazing cell type has so far defied a generalized classification system. Interneuron complexity within the telencephalon could be simplified by viewing them as elaborations of a much more finite group of developmentally specified cardinal classes that become further specialized as they mature. Our perspective emphasizes that the ultimate goal is to dispense with classification criteria and directly define interneuron types by function.

Interneurons, of all the cells within the forebrain, are the most diverse in terms of morphology, connectivity and physiological properties[1]. Until 10 years ago, their classification, with a few notable exceptions[2], remained descriptive. Moreover, interneuron diversity was often treated either as a quasi continuum or a diversity space with cell types numbering potentially in the hundreds[3,4]. Studies from the past few years have coalesced into the surprising view that interneuron diversity may fundamentally be far more limited. When we consider their commonalities at a genetic, circuit or functional level, an argument can be made for condensing large subclasses of interneurons into more finite groups. In this Review, we suggest that, on the basis of both developmental and functional criteria, interneuron diversity can be simplified and addressed experimentally. The differences in connectivity, gene expression and physiological properties of interneurons found across the brain seem enormous (Fig. 1), nevertheless, we argue that this complexity arises from a small number of non-overlapping cardinal classes. These represent developmental genetic ground states that can further specialize through their later interactions with other neurons. The ultimate goal of defining their identity through a set of computational principles remains daunting. However, with the advent of new tools that provide unprecedented targeting specificity, coupled with the means to manipulate the *in vivo* activity of targeted neural populations this goal is becoming attainable.

Birth and specification of interneurons

How is neuronal diversity created? Developmental studies across various species[5,6] and systems[7,8] have suggested that cell diversity arises from specification programs established in progenitors that have been modified to varying extents by their subsequent post-mitotic interactions. The balance between genetic- and experience-dependent processes seems to represent a compromise dictated by the organizational constraints of the particular system. Within the cortex, pyramidal cells undergo a relatively orderly migration from the proliferative zone to the overlying cortical plate. As such, cell identities are largely controlled by programs established within progenitors[9]. By contrast, interneuron progenitors of the telencephalon undergo incredibly complex patterns of dispersion. At the extremes, this could either be due to exquisitely precise pre-programs for migration to particular structures or plasticity mechanisms that allow the progenitors to adapt to local environments.

Until the late nineties, it was widely assumed that excitatory and inhibitory neurons within the cortex shared a common lineage. The seminal breakthrough came from the realization that interneurons originated within focal subcortical proliferative zones[10]. This first came to light with landmark papers showing that the GABAergic populations from the ganglionic eminences migrated dorsally to populate the cortex[10], as well as to all other structures within the telencephalon[11,12]. Subsequent work in the spinal cord, led to the conjecture that an understanding of how specific subtypes are generated would come from a detailed analysis of gene expression within progenitors. It was assumed that combinatorial transcriptional codes in subpallial progentiors functioned to establish distinct cortical interneuron subtypes.

The connection between developmental origins and interneuron diversity has steadily expanded over the past 20 years. Almost all GABAergic interneurons within the telencephalon arise from one of two embryonic subcortical progenitor zones, the medial ganglionic eminence (MGE) and caudal ganglionic eminence (CGE; Fig. 2). Moreover, those arising from each structure represent complementary interneuron subtypes[6,11–15]. These major areas are augmented by specialized subpopulations from the lateral ganglionic eminence[9] and the preoptic region[8]. It also became clear that there is a strong correspondence between interneuron classes and the specific progenitor zones that gives rise to them. Within the cortex, the MGE gives rise to the parvalbumin (PV)-expressing fast-spiking interneurons (including both basket and chandelier cells) and the somatostatin (SST)-expressing populations, of which the Martinotti cells form the largest subset[13,16,17]. The CGE produces relatively rarer subtypes, including neurogliaform, bipolar and vasointestinal peptide (VIP)-expressing multipolar interneurons[14].

Genetic lineage analysis within the hippocampus reinforces the idea that specific interneurons arise from specific structures but demonstrates that a simple correspondence across forebrain regions is untenable. For instance, although in the cortex neurogliaform neurons are CGE-derived, a large proportion of the corresponding population in the hippocampus arises from the MGE[18]. Furthermore, whereas some classes, such as fast-spiking basket cells, show marked similarities across structures, others subclasses do not yet seem to have obvious paralogues. For instance, cholecystokinin basket cells, although a large population within the cortex and hippocampus, do not seem to be present within other brain areas[19]. Similarly, there seems to be at least two populations of the so called oriens/lacunosum-moleculare cells (named in accordance with the position of their cell body and dendrites[20,21]) that derive from distinct sources, one which expresses the ionotropic serotonin receptor 5HT3aR and one that does not. Adding further complexity,

[1]Cold Spring Harbor Laboratory, Marks Building, New York 11724, USA. [2]NYU Langone Medical Center, First Avenue, Smilow Research Building, New York 10016, USA.

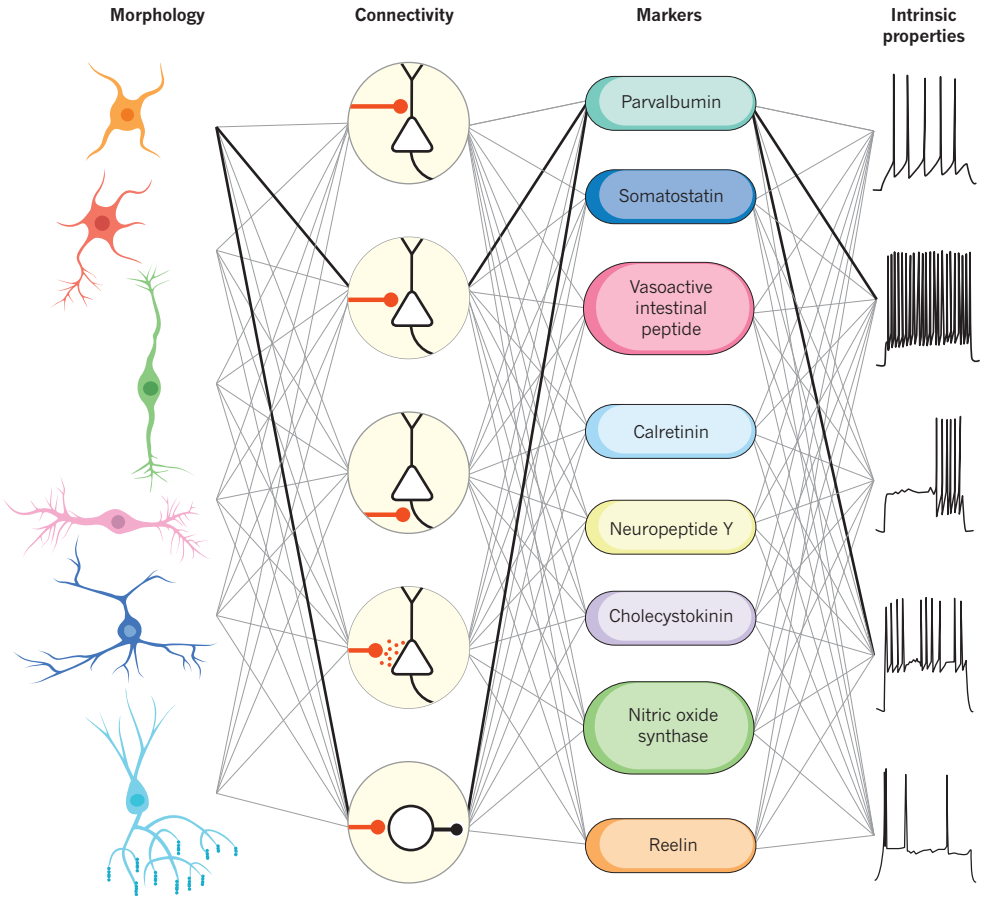


Figure 1 | Multiple dimensions of interneuron diversity. Interneuron cell types are usually defined using a combination of criteria based on morphology, connectivity pattern, synaptic properties, marker expression and intrinsic firing properties. The highlighted connections define fast-spiking cortical basket cells.

analysis of the basal ganglia suggests that only the MGE is a major source of interneuron populations within these structures[22].

These differences across areas raise two possibilities. First, there might be dedicated populations of interneuron progenitors that are committed to populating specific brain structures. Second, the notion of referring to an interneuron's origin as deriving from a specific embryonic structure may be an imprecise proxy for gene expression. For instance, even though hippocampal neurogliaform cells arise from both the MGE and CGE, a common constellation of specification genes may be acting within both embryonic regions. Similarly, the differential expression of functional determinants such as serotonin receptors, in otherwise similar interneuron subtypes, are unlikely to represent distinct cardinal classes. Rather they probably represent iterations produced by cardinal cousins or differential post-mitotic interactions by members of a single cardinal class.

These details emphasize the importance of mapping interneuron diversity onto molecular mechanisms. GABAergic lineages can be divided into those with long-range projections, such as those in the striatum or globus pallidus, and interneuron populations that largely project locally. A number of factors seem to be used within all GABAergic neurons (Fig. 2), most notably the transcription factors encoded by *Dlx1* and *Dlx2*, *Ascl2*, and *Gsx1* and *Gsx2* that themselves form a regulatory network[23–25]. In the pallium (the region of the forebrain that will give rise to cortical structures), a similar cohort of transcription factors, including those encoded by *Emx1*, *Neurog1* and *Neurog2*, and *Pax6*, function analogously in the specification of the excitatory populations[26].

Dlx1 and *Dlx2*, in particular, function at multiple stages of GABAergic maturation: in the acquisition of GABAergic identity[27], the initiation and cessation of tangential migration[4,28,29], and in the morphological and physiological maturation of specific subclasses[29]. The specific role of *Dlx1* and *Dlx2* in these disparate developmental activities has become clearer as their transcription targets have been identified. These include *Elmo1*, *Dlx5* and *Dlx6*, *Arx* and *Gemin2* (or *Zep2*), each of which has been shown to be required in the control of migration and regional identity[30–33]. Moreover, mutations of these genes, presumably through their requirement for interneuron function, contribute to a variety of affective psychiatric disorders[30].

In addition, a number of factors seem to be more restricted to specific subtypes. Although far from complete, a genetic hierarchy for the MGE-derived PV and SST lineages has begun to emerge. Within the MGE, the cascade begins with *Nkx2-1*, which acts as master regulator in promoting MGE-derived interneuron fates over CGE-derived cell types[34,35]. Moreover, in the clearest example of a single gene contributing to the generation of a specific interneuron subtype, chandelier or axo-axonic interneurons have been shown to arise relatively late in embryonic development (embryonic day 15 to 18 in mice) from a population of *Nkx2-1* progenitors[36,37]. In addition, *Nkx2-1* is a gene with both activator and repressor function. Its repressor function attenuates the expression of CGE-specific genes, whereas its activator function induces the expression of Lhx6 (ref. 17), which is needed to promote the differentiation of both PV- and SST-expressing interneurons. Lhx6 in turn drives the expression of a series of factors including *Sox6* and *Satb1* (refs 38, 39) the products of which selectively affect the development of both PV- and SST-expressing interneurons. By contrast, so far only a few genes have proven to be specific to the CGE-derived lineages. Collectively, the specific functions of these transcription factors and their targets are still a work in progress, but the tools to tackle this challenge are at hand.

Although the emerging picture is exhilarating, rather than coalescing into an explanation for the myriad distinct subtypes that populate all areas of the forebrain (Fig. 1), it seems to only reveal a handful of genetic cascades. If the goal were simply to account for the neuronal markers

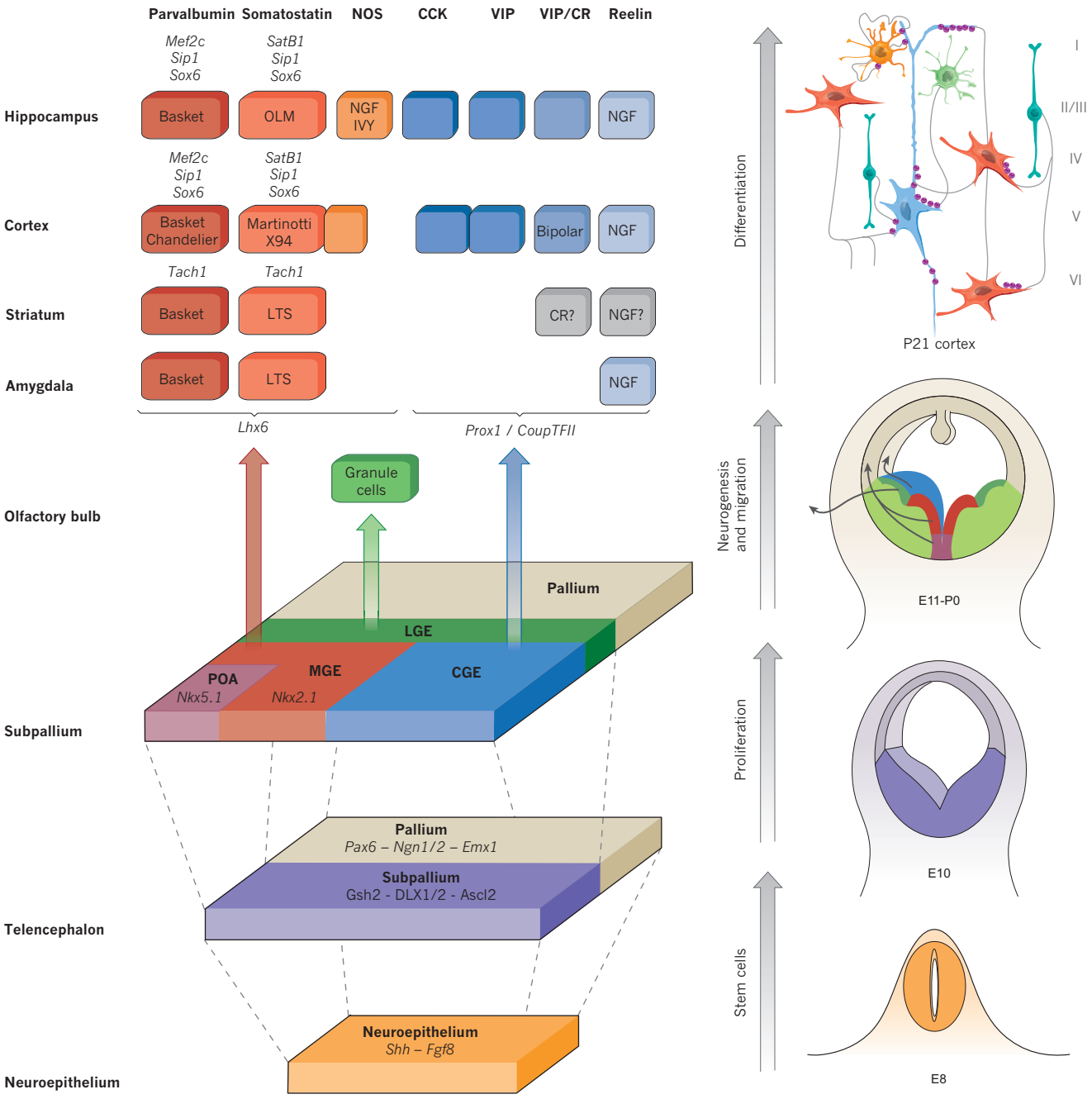


Figure 2 | Interneuron subtypes are generated from discrete proliferative regions within the subpallium. The progressive development of the telencephalon from an undifferentiated epithelium into discrete proliferative zones that produce particular interneuron populations and the specific genes involved at each stage. On the right are more anatomically accurate cross-sections of the progenitor zones from embryonic day 8 to 11, and a schematic of interneuron diversity in the cortex. Interestingly, although common proliferative zones produce the entire diversity of interneurons across all telencephalic structures, unique cell types and gene expression are seen in interneuron populations that reside in particular telencephalic structures. CCK, cholecystokinin; CR, calretinin; CGE, caudal ganglionic eminence; LGE, lateral ganglionic eminence; LTS, low threshold spike; MGE, medial ganglionic eminence; NGF, nerve growth factor; NOS, nitric oxide synthase; OLM, oriens/lacunosum moleculare; POA, pre-optic area; VIP, vasointestinal peptide.

that have conventionally been used to categorize interneurons, then a mere six (PV, SST, VIP, nitric oxide synthase, reelin and calretinin) could divide most interneurons within the forebrain. But clearly such a classification would belie the regional complexity of interneurons. Within the hippocampus alone there are easily four or five SST-expressing cell types and at least three PV-expressing populations. Similar distinct subpopulations of SST- and PV-expressing populations have been discovered in the cortex, and more will probably be found. Although we believe that the cardinal specification of interneurons is only the first, albeit crucial, step in the progressive specification of subpallial progenitors, can our cardinal identity hypothesis account for this increasing wealth of interneuron subtypes? It is certainly possible that we have grossly underestimated the cardinal subtypes. Complex maps showing intricate embryonic patterns of gene expression within the subpallium have been posited to combinatorially specify different cell types[40]. However, we think that the cell types generated by developmental programs are unlikely to explain all of the regional diversity observed. First, the loss of specific genes results in phenotypes that are invariably not restricted to specific interneuron subtypes. Second, the loss of specific genes affects the generation of interneurons across various structures, arguing against

the existence of progenitor populations dedicated to the generation-specific interneuron classes. That said, it is possible to imagine a combinatorial gene regulation strategy in which an individual gene could be necessary for a variety of disparate differentiation programs. Hence, another way to explore the question of whether regional diversity is established in progenitors is through lineage analysis.

Interneuron lineages

Are the interneuron populations that populate particular structures, such as the hippocampus or cortex, derived from dedicated progenitor pools? Two recent studies have directly explored the role of lineage in the development of cortical interneurons[41,42]. Both have shown that clonally related progenitors seem to be preferentially relegated to specific cortical columns or layers, supporting the idea that progenitor lineages are dedicated to producing lineages destined for particular brain structures. Given the long and convoluted paths taken by these progenitors as they transit to their mature position[43,44], it was stunning that clones could collectively target particular parts of the cortex. Interestingly, such clones were equally likely to be comprised of mixed SST- and PV-expressing interneurons rather than one subtype. Perhaps this lack of tendency for clones to 'breed true' should not come as a surprise. A wealth of lineage analysis in invertebrates[45], as well as in the vertebrate retina[31] and spinal cord[32], indicates that neuronal lineages, although stereotyped, do not generally produce cells of a single subtype. Moreover, it will be interesting to explore whether in addition to being clustered, lineally related cells are also dispersed and, if so, to what extent[33,46]. If they are dispersed, it will be intriguing to assess the afferent and efferent connectivity of such clones, as this would help address the question: to what extent do intrinsic compared with local cues direct the connectivity of interneurons?

How circuits nurture interneuron subtypes

The accumulated evidence supports a strong role for developmentally regulated genetic programs in the allocation of interneurons to broad cardinal classes. What it does not seem to explain is how interneurons from the same cardinal class are able to form connections with such a wide variety of synaptic partners. In favour of a role for local cues contributing to this process, various studies have suggested that both excitatory and inhibitory signals may influence the migration and positioning of developing interneurons[47,48]. Recent data have shown that attenuating the activity of specific interneuron populations affects both their migration and morphological development[49]. Although the activity is class specific, whether it is instructive rather than simply permissive has yet to be demonstrated. That said, there seems to be growing support for the notion that local signals may direct the region-specific differentiation of interneurons. In support of this, a number of genes are specific to interneurons within the cortex but not the striatum, including *Zep2*, *Dlx1*, *Elmo1* and *Mef2c*, the last three of which are activity regulated[30,49–52]. Although it may be that activity simply promotes the maturation of interneuron populations that are already pre-specified, it is also possible that activity directs region-specific differentiation. Extensive work has indicated that voltage-gated calcium influx may result in *de novo* gene expression (reviewed in ref. 53). It is the failure at present to identify genes within proliferative populations that are indicative of region-specific differentiation that has led us to propose a two-phase model of interneuron specification. We envision that activity-regulated gene expression during critical periods may be responsible, during the second phase, for the allocation of cardinal classes into specific subclasses. Recent work has shown that MGE-derived cells can productively integrate into both normal and abnormal neuronal circuits[54–56]. This supports the idea that local cues can direct nascent interneurons to form appropriate connectivity with various synaptic partners. How interneurons form functional circuits in a variety of structures is a crucial question that remains to be answered.

So far we have taken a bottom-up developmental view that is aimed at examining events by which interneuron subtypes are integrated into functional circuits. In the next section, we take a top-down view and examine circuit-specific functions of interneuron types. A prediction of our model is that the developmental genetic programs functioning in interneuron progenitors lead to the production of a relatively small number of cardinal subclasses. We believe that the much larger diversity of interneurons observed in mature brain circuits reflects later refinements imposed locally on specific subclasses. If this were true, it would predict that interneurons from the same cardinal class would, within the same circuit, be exposed to similar cues and hence develop similar functional properties. Although the data so far are in nascent stages, the availability of genetic driver lines[57,58] to reliably target particular interneuron cohorts, has provided the means to test this hypothesis.

How interneurons function

What is a meaningful measure of the function of an interneuron subtype? Because interneurons generally project locally, their firing needs to be understood in the context of the circuits to which they contribute. There are two complementary ways of approaching the question of what an interneuron does: by examining when and how they are recruited to fire or by determining the impact of their firing on the circuit (Fig. 3). First, understanding recruitment requires us to determine under what circuit and brain-state configurations or behavioural contingencies is a given neuron active? This is strongly constrained by a neuron's afferent connectivity, which in turn is probably dictated by its developmental genetic program. However, interneurons are not hardwired. A set of afferents that drive interneurons to fire in one context may fail to do so in another. Clearly, to understand their recruitment we must take into account a large number of factors. The classic idea is that interneurons

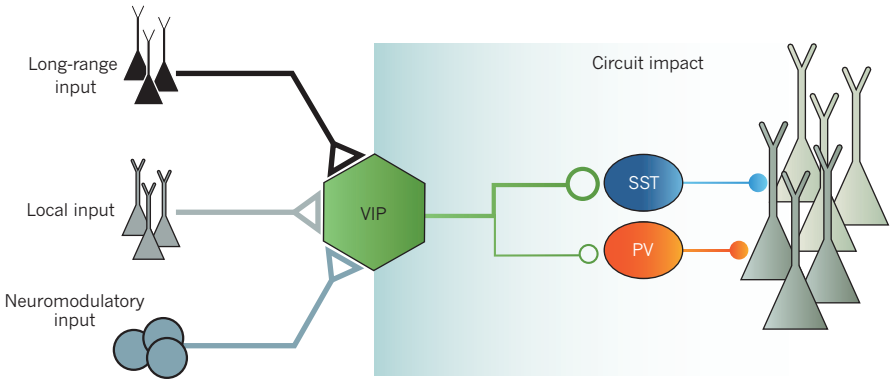


Figure 3 | Two faces of interneuron function. A cortical circuit from the perspective of a vasointestinal peptide (VIP) interneuron. The recruitment of VIP interneurons is constrained by the inputs they receive. The afferents can be local, long range from other cortical areas as well as neuromodulatory from the dorsal raphe and nucleus basalis through ionotropic receptors. The circuit impact of VIP interneurons is constrained by their outputs. The efferents are mostly to somatostatin (SST) interneurons and to a smaller degree to parvalbumin (PV) interneurons, which lead to the disinhibition of a functional subset of principal cells[85,101,102].

coordinate networks, such that their recruitment is best understood in reference to the local population activity (Fig. 4a). We also consider an alternative view that we term the flow-control hypothesis, whereby interneurons gate information flow within a given circuit and are excited at precise moments in reference to specific behavioural events (Fig. 4b). These two ideas are by no means mutually exclusive, as successful flow control must depend on the signals being suitably coordinated. Second, we can ask about circuit impact. How does the firing of an interneuron influence the activity of neurons in its local circuit? This aspect of function is strongly constrained by the efferent connectivity of a neuron, which is thought to be dependent on its developmental genetic program. In addition, the impact of interneuron type will also greatly depend on whether their recruitment is coordinated with other neurons of the same cohort.

Historically, models of cortical function have focused on circuit motifs, repeated patterns of connectivity, to infer computational function for specific cell types[59]. Perhaps because inhibitory interneurons are largely local, they have generally been considered to simply guard excitatory networks against runaway excitation[60]. Recently, our understanding of their function has become significantly more sophisticated. Among other lessons, we learned that interneurons could normalize the activity of local excitatory networks as well as provide feedforward inhibition. The latter strongly influences the timing of signals and allows excitatory signals to remain sub-threshold while carrying information. Of course, these are just two motifs out of a vast range of possibilities, including cross-coupling that can lead to synchronization, lateral inhibition that can segregate principal neuron populations, and disinhibition that can generate elevated activity. It has become clear that there are at least as many inhibitory circuit motifs as there are cell types.

Computing with interneurons

How then does the diversity of interneurons contribute to neural computations? First, it is worth noting that it would be difficult to imagine networks with only excitation. In fact, from an engineering standpoint, such networks would have to have extremely time-limited dynamics or they would become intrinsically unstable. Moreover, inhibition not only provides balance, it also ensures richness in the possible dynamics within networks of principal neurons. These considerations led to the idea that interneuron diversity allows for a vast increase in the computational power of cortical circuits[61,62]. Broadly speaking, the computational functions of interneurons can be grouped into either arithmetic or timing.

A long-held idea is that different interneurons perform essentially arithmetic operations, such as subtraction or division[63,64]. Inhibition can provide gain control by changing the input–output relationship between the excitatory drive and the resulting firing rate in principal cells: either by decreasing its slope divisively or by a subtractive shift. In turn, these elementary operations are the building blocks for cortical computations such as normalization, an operation that provides divisive gain in proportion to the summed activity in a circuit[65,66]. Such gain modulation might result from shunting[67], synchronous[68] or balanced[69] inhibition. Originally proposed to explain early visual cortical responses[65], normalization has become one of the most studied cortical computations[70,71].

Another line of theoretical investigation has focused on the role of interneurons in controlling the timing of neural activity. More complex network functions require that neurons do not fire together. This can be achieved by dynamically balancing excitation with inhibition so that the resulting network activity becomes temporally irregular and asynchronous. These balanced networks thus provide rich dynamics and rapid responses[71]. Indeed, cortical recordings often reveal finely balanced excitation and inhibition[72–75], consistent with these models. When inhibition precisely tracks excitation, it can also increase temporal precision[72,76] or decorrelate networks[77].

Circuit impact of identified interneuron types

To understand how computations are implemented in neural networks requires an appreciation of how distinct interneuron subtypes affect local networks. Although conventionally studied *in vitro*, we focus mostly on recent *in vivo* work using transgenic mice for targeting interneurons on the basis of markers, such as PV, SST and VIP[57,58]. It should be noted that these Cre-driver mouse lines neither demarcate entirely homogeneous interneuron populations nor map precisely onto cardinal interneuron types. Nevertheless, they provide a convenient and powerful tool for parsing interneuron heterogeneity because these three major markers delineate distinct non-overlapping populations and in aggregate can label around 85% of all cortical interneurons[13,78,79]. When combined with optogenetic modulators[80,81], they have allowed researchers, for the first time, to test many long-held theories about the roles of inhibition.

PV-expressing interneurons (either soma-targeting basket cells or chandelier cells targeting the axon initial segment) are strategically positioned to control spiking, and are also strongly interconnected, which promotes their synchronous activity[82–86]. Recent studies have shown that this population controls the timing of spikes with respect to theta oscillations in the hippocampus[87–89]. In the visual cortex, optogenetic control

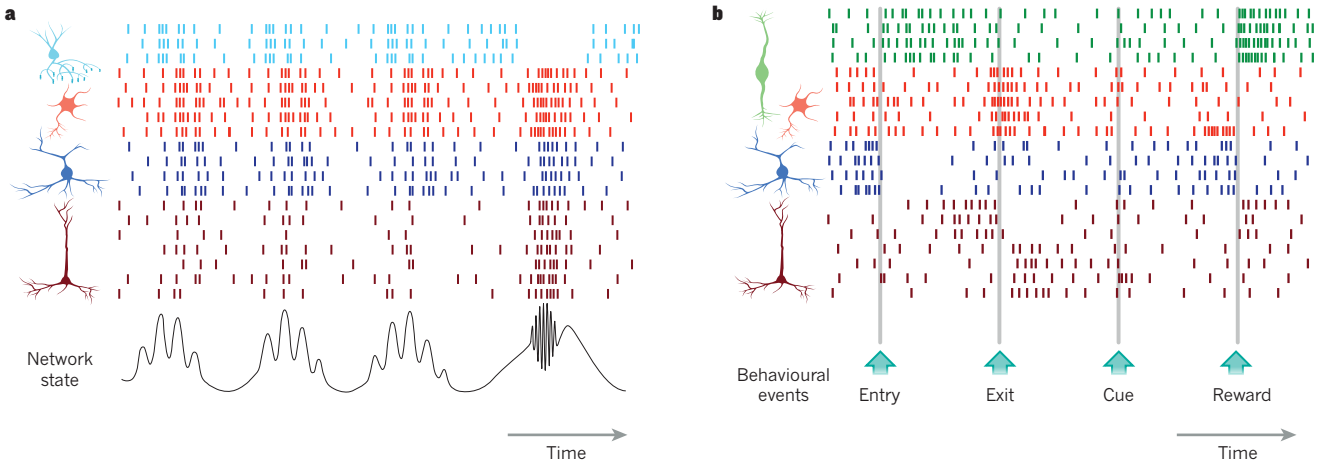


Figure 4 | Coordination and flow control hypotheses of recruitment. **a**, Coordination hypothesis. The bottom trace shows a local field potential representing the network state in the hippocampus. The firing of different neuron types (chandelier cell, light blue; basket cell, red; OLM cell, blue; pyramidal cell, brown) can be described in reference to the local field potential, both in terms of overall activity level and phase relationship[87,107,113]. **b**, Flow control hypothesis. The bottom arrows mark the timing of four behavioural events: entry, exit, cue and reward. The firing of different neuron types (vasointestinal peptide, green; parvalbumin, red; somatostatin, blue; pyramidal cell brown) can be described in reference to these events[84,102].

of PV-expressing interneurons can bidirectionally modulate the gain of visual responses[90,91]. Under some conditions, optogenetic activation of this population can even sharpen the tuning of cortical responses[92].

SST-expressing cortical Martinotti neurons are dendritic-targeting interneurons that project to layer 1 of the cerebral cortex and provide inhibition to the tufts of deep-layer pyramidal cells. Inhibition by Martinotti cells strongly suppresses dendritic calcium spikes and bursting[93], and can mediate di-synaptic inhibition between neighbouring pyramidal cells[94,95]. Similarly, in the hippocampus, dendritic inhibition by SST neurons controls burst firing[87,88,96]. In the cortex, SST-expressing interneurons have broader spatial tuning and can mediate surround suppression of visual responses[70,97]. However, SST-expressing neurons are anatomically diverse, with some subtypes specializing in disinhibition of local principal cells[98].

VIP demarcates the third major class of interneurons, which constitute around 15% of all interneurons[78,79] and are mostly located in the superficial layers of cortex. VIP-expressing interneurons have long been proposed to mediate disinhibition[99,100]. Recent studies have shown that in four different cortical regions, VIP interneurons tend to inhibit most SST and a smaller fraction of PV-expressing interneurons[85,101,102]. Inhibition of these interneurons in turn disinhibits principle cells *in vivo,* providing a form of gain control[102]. These results demonstrate that VIP-expressing neurons form a disinhibitory microcircuit that is conserved across cortical regions with shared computational functions (Fig. 3).

A final example is provided by layer 1 neurons, which are almost all inhibitory interneurons[103]. Recent results have shown that two major interneuron types have opposing functions: neurogliaform cells inhibit layer 2/3 pyramids, whereas single bouquet cells inhibit other interneurons within layer 2/3, and may provide disinhibition *in vivo*[104,105].

These studies injected great excitement into the field as they provided, at last, the causal testing of hypotheses using optogenetics. In addition, they provided support for the hypothesis that cardinal classes of interneurons have defined circuit functions. However, they carry the caveat that they assume that these populations are synchronously activated under physiological conditions. Although often this is a reasonable assumption, it needs to be tested on a case-by-case basis by recording the requisite neural population during behaviour.

Coordination and flow control during recruitment

What are the brain-state and behaviour-dependent contingencies that determine when a specific interneuron type is activated? At the broadest scale, different behavioural modes are associated with large changes in global brain activity, therefore it is not surprising that different classes of inhibitory interneurons are activated in a highly state-dependent manner[106–109]. At a more refined scale, what is the simplest description of the conditions under which a given interneuron is activated? Is recruitment of an interneuron subtype best understood with reference to a network state or a behavioural contingency?

Network recruitment of interneurons

One idea is that interneurons coordinate the precise timing of principal cell activation, such as network oscillations. There is a long history of experimental and theoretical investigations proposing that the diversity of interneuronal subtypes underlies a division of labour for organizing cortical population activity at different time scales[61,110–112].

The best-studied examples of the contribution of different interneurons come from the hippocampus, in which recordings from targeted cells have been used to correlate their firing to network oscillations[107,113–116]. These studies, mostly in anaesthetized animals, revealed that distinct interneuron subtypes fire during different rhythms (for example, theta, gamma and ripple) and with distinct phase relationships, suggesting that they differentially contribute to network dynamics. Thus it seems that the spike timing of different interneuron types can be referenced to specific network events[61,107,117] (Fig. 4a). More generally, this work supports the coordination hypothesis, whereby each interneuron subtype performs as a temporal specialist within a 'distributed clock system' that coordinates pyramidal cell ensembles[112].

These observations set the stage for testing the causal role of different interneuron subtypes in generating oscillations[106,107]. Recently, two studies probed the role of PV interneurons using either optogenetic activation[118] or suppression[119], and found increased and decreased gamma oscillations, respectively, suggesting that interneurons are indeed actively involved in their generation.

Behavioural recruitment of interneurons

An alternative view is that the function of some interneuron types is better described with reference to behavioural events. Although interneuronal identity has long been inferred in behaving animals on the basis of spike waveform and firing pattern[110,120,121], it is only recently that this could be directly ascertained for a handful of genetically defined interneurons during well-controlled behaviours. For instance, a recent study in the anterior cingulate cortex of mice reported the surprising observation that deep-layer PV-expressing and narrow-spiking SST-expressing interneurons responded in a functionally homogeneous manner at specific behavioural epochs[84] (Fig. 4b). Similar observations were made in rat motor cortex using juxtacellular labelling in head-fixed rats: pyramidal cells responded heterogeneously, whereas all PV interneurons responded similarly at the moment of movement initiation[122]. This shows that PV interneurons, at least in mouse frontal regions, can be thought of as a functional unit. How could such functional homogeneity be achieved? One possibility is that inhibitory interneurons strongly sample local principal cell activity and their activation reflects a summary of local activity[82]. Therefore PV interneurons might fire in a behaviour- and region-dependent manner, which may be the 'leaving decision' in the anterior cingulate cortex, owing to its role in foraging behaviours[123], but movement initiation in the motor cortex.

Other examples of behaviourally activated responses come from the auditory cortex, in which a large fraction of interneurons in layer 1 are activated by negative reinforcers during auditory fear conditioning[105]. Similarly, VIP interneurons were observed to be strongly and uniformly recruited by negative- (air puff or mild shock) or positive- (water reward) reinforcement during an auditory discrimination task[102] (Fig. 4b). Although such reinforcement feedback-related signals may at first seem surprising in a primary sensory area, as VIP interneurons mediate disinhibition (discussed earlier) they are an ideal conduit for gating signals[99] (Fig. 3).

As we stated earlier, because interneurons are embedded in a highly interconnected network, their functions need to be understood in the context of local networks[124]. In light of this, the observations that some interneuron types are recruited at specific behavioural events may seem puzzling. Indeed, consistent with the coordination hypothesis[112,117], one would expect that most responses would be constrained by the state of the local network, on a time scale of milliseconds, and not by behavioural contingencies. Mechanistically, the observed behaviourally specific activation may reflect local network activity, which itself is tied to specific behavioural contingencies. We suspect that this explains the homogeneous activation of deep-layer PV neurons. Nevertheless, their function is most parsimoniously described by temporal reference to specific behavioural events. Alternatively, some classes of interneurons may be activated by strong long-range inputs. For instance, neuromodulatory systems can provide behaviour-dependent inputs to specific interneuron classes. Interestingly, VIP neurons have ionotropic receptors for the neuromodulators acetylcholine and serotonin, which probably drive reinforcement signals in these neurons[99,125].

At present, the behavioural recruitment of many interneuron types remains unexplored and different mechanisms may apply to each type. In contrast to the coordination hypothesis, some early results support the flow-control hypothesis (Fig. 4), proposing that distinct interneuron types specialize in controlling information flow in and out of local circuits during specific behavioural contingencies, and thus acting much like controllers of a state machine. This suggests that the recruitment

of an interneuron type is linked with behavioural scale requirements of local circuits. Moreover, the observations that genetically defined interneuron classes show similar recruitment suggests they do act as functional units, supporting the existence of a small number of cardinal interneuron types.

Outlook

We are in the midst of an exciting era in which each week new data on the development and function of interneurons are being brought to light. In this Review, we suggest that the large diversity in interneuron classes may originate from a handful of cardinal cell types. Such an assertion could be misinterpreted as a statement claiming that interneuron classes are not diverse or that divisions into further subtypes are not warranted. The incredible work in areas such as the hippocampus shows us that this is patently incorrect. We have provided a framework that we hope will help to direct future studies by consolidating interneuron diversity into cardinal classes with specific ground states. Therefore if one wishes to explore questions regarding intrinsic physiological properties, axonal targeting or general target selection, understanding the ground state established in progenitors is a good place to start. However, to explore the circuit properties, connectivity or computational contributions of a subclass of interneurons, one needs to consider the interplay between cardinal cells and the local cues received from the circuits that they contribute to. From a functional point of view, if we confine ourselves to specific circuits, such as VIP interneurons, a cardinal class will share important aspects of their function. Hence the tools to genetically target cardinal classes will prove invaluable for parsing the function of the different and more exotic interneuron subtypes they ultimately give rise to.

It is intriguing to contemplate why such a mechanism is used to create cellular diversity. It may be that the strategy used by natural selection favours simple programs to provide stability, and combinatorial assembly to provide complexity. Although they are ultimately incorporated in a wide breadth of circuits, cardinal interneuron classes share crucial combinations of features that enable their function. Genetic programs direct the receptors they express, the cell types and subcellular compartment they innervate, as well as their firing properties. These features in turn strongly constrain interneuron recruitment and circuit impact. In short, we may ultimately find that interneurons exist as cardinal classes because nature has conspired to bestow on them generalized computational function that necessitated the presence of common biophysical and hodological properties. Despite these commonalities, it is self-evident that neural circuits allow for a remarkable array of behavioural outcomes. Harnessing biology's ability to use a limited set of building blocks, to create enormous diversity circuits holds the real promise that we may soon begin to understand the means by which brain circuits self-assemble and initiate function. ■

Received 18 September; accepted 25 November 2013.

1. Ascoli, G. A. *et al*. Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nature Rev. Neurosci*. **9,** 557–568 (2008). **This is the best effort to date by physiologists, anatomists and developmental neurobiologists to come to a common nomenclature for GABAergic interneurons.**
2. Freund, T. F. & Buzsáki, G. Interneurons of the hippocampus. *Hippocampus* **6,** 347–470 (1996).
3. Markram, H. *et al*. Interneurons of the neocortical inhibitory system. *Nature Rev. Neurosci*. **5,** 793–807 (2004).
4. Parra, P., Gulyás, A. I. & Miles, R. How many subtypes of inhibitory cells in the hippocampus? *Neuron* **20,** 983–993 (1998).
5. Brody, T. & Odenwald, W. F. Regulation of temporal identities during *Drosophila* neuroblast lineage development. *Curr. Opin. Cell Biol.* **17,** 672–675 (2005).
6. Xu, Q., Cobos, I., De La Cruz, E., Rubenstein, J. L. & Anderson, S. A. Origins of cortical interneuron subtypes. *J. Neurosci.* **24,** 2612–2622 (2004).
7. Leone, D. P., Srinivasan, K., Chen, B., Alcamo, E. & McConnell, S. K. The determination of projection neuron identity in the developing cerebral cortex. *Curr. Opin. Neurobiol.* **18,** 28–35 (2008).
8. Gelman, D., Griveau, A. & Dehorter, N. A wide diversity of cortical GABAergic interneurons derives from the embryonic preoptic area. *J. Neurosci.* **31,** 16570–16580 (2011).
9. O'Leary, D. D. M. & Borngasser, D. Cortical ventricular zone progenitors and their progeny maintain spatial relationships and radial patterning during preplate development indicating an early protomap. *Cereb. Cortex* **16,** i46–i56 (2006).
10. Anderson, S. A., Eisenstat, D. D., Shi, L. & Rubenstein, J. L. Interneuron migration from basal forebrain to neocortex: dependence on *Dlx* genes. *Science* **278,** 474–476 (1997). **This paper demonstrates that cortical interneurons are derived subpallially.**
11. Wichterle, H., Turnbull, D. H., Nery, S., Fishell, G. & Alvarez-Buylla, A. *In utero* fate mapping reveals distinct migratory pathways and fates of neurons born in the mammalian basal forebrain. *Development* **128,** 3759–3771 (2001). **This paper provided *in vivo* evidence that specific interneurons are derived from specific embryonic progenitor zones.**
12. Nery, S., Fishell, G. & Corbin, J. G. The caudal ganglionic eminence is a source of distinct cortical and subcortical cell populations. *Nature Neurosci*. **5,** 1279–1287 (2002).
13. Butt, S. J. *et al.* The temporal and spatial origins of cortical interneurons predict their physiological subtype. *Neuron* **48,** 591–604 (2005).
14. Lee, S., Hjerling-Leffler, J., Zagha, E., Fishell, G. & Rudy, B. The largest group of superficial neocortical GABAergic interneurons expresses ionotropic serotonin receptors. *J. Neurosci*. **30,** 16796–16808 (2010).
15. Miyoshi, G. *et al*. Genetic fate mapping reveals that the caudal ganglionic eminence produces a large and diverse population of superficial cortical interneurons. *J. Neurosci*. **30,** 1582–1594 (2010).
16. Fogarty, M. *et al.* Spatial genetic patterning of the embryonic neuroepithelium generates GABAergic interneuron diversity in the adult cortex. *J. Neurosci*. **27,** 10935–10946 (2007).
17. Du, T., Xu, Q., Ocbina, P. J. & Anderson, S. A. NKX2.1 specifies cortical interneuron fate by activating Lhx6. *Development* **135,** 1559–1567 (2008).
18. Tricoire, L. *et al*. Common origins of hippocampal Ivy and nitric oxide synthase expressing neurogliaform cells. *J. Neurosci.* **30,** 2165–2176 (2010). **This paper compared references 11–15 to demonstrate that similar interneuron subtypes in the cortex versus the hippocampus could be derived from distinct progenitor zones.**
19. Armstrong, C. & Soltesz, I. Basket cell dichotomy in microcircuit function. *J. Physiol.* **590,** 683–694 (2012).
20. McBain, C. J. & Fisahn, A. E. Interneurons unbound. *Nature Rev. Neurosci.* **2,** 11–23 (2001).
21. Chittajallu, R. *et al*. Dual origins of functionally distinct O-LM interneurons revealed by differential 5–HT3AR expression. *Nature Neurosci*. **16,** 1598–1607 (2013).
22. Marín, O., Anderson, S. A. & Rubenstein, J. L. R. Origin and molecular specification of striatal interneurons. *J. Neurosci.* **20,** 6063–6076 (2000). **This paper reports our best understanding so far of the origins of striatal interneurons.**
23. Wang, B., Waclaw, R. R., Allen, Z. J., Guillemot, F. & Campbell, K. Ascl1 is a required downstream effector of *Gsx* gene function in the embryonic mouse telencephalon. *Neural Dev*. **4,** 5 (2009).
24. Wang, B. *et al*. Loss of Gsx1 and Gsx2 function rescues distinct phenotypes in Dlx1/2 mutants. *J. Comp. Neurol*. **521,** 1561–1584 (2013).
25. Long, J. E., Cobos, I., Potter, G. B. & Rubenstein, J. L. Dlx1&2 and Mash1 transcription factors control MGE and CGE patterning and differentiation through parallel and overlapping pathways. *Cereb. Cortex* **19** (Suppl. 1), i96–i106 (2009).
26. Schuurmans, C. & Guillemot, F. Molecular mechanisms underlying cell fate specification in the developing telencephalon. *Curr. Opin. Neurobiol*. **12,** 26–34 (2002).
27. Stühmer, T., Anderson, S. A., Ekker, M. & Rubenstein, J. L. R. Ectopic expression of the *Dlx* genes induces glutamic acid decarboxylase and Dlx expression. *Development* **129,** 245–252 (2002).
28. Cobos, I., Borello, U. & Rubenstein, J. L. Dlx transcription factors promote migration through repression of axon and dendrite growth. *Neuron* **54,** 873–888 (2007).
29. Cobos, I. *et al.* Mice lacking Dlx1 show subtype-specific loss of interneurons, reduced inhibition and epilepsy. *Nature Neurosci.* **8,** 1059–1068 (2005).
30. Colombo, E. *et al*. Inactivation of *Arx*, the murine ortholog of the X-linked lissencephaly with ambiguous genitalia gene, leads to severe disorganization of the ventral telencephalon with impaired neuronal migration and differentiation. *J. Neurosci.* **27,** 4786–4798 (2007).
31. Bassett, E. A. & Wallace, V. A. Cell fate determination in the vertebrate retina. *Trends Neurosci*. **35,** 565–573 (2012).
32. Leber, S. M., Breedlove, S. M. & Sanes, J. R. Lineage, arrangement, and death of clonally related motoneurons in chick spinal cord. *J. Neurosci.* **10,** 2451–2462 (1990).
33. Walsh, C. & Cepko, C. L. Clonally related cortical cells show several migration patterns. *Science* **241,** 1342–1345 (1988).
34. Sussel, L., Marín, O., Kimura, S. & Rubenstein, J. L. Loss of *Nkx2.1* homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: evidence for a transformation of the pallidum into the striatum. *Development* **126,** 3359–3370 (1999).
35. Butt, S. J. *et al.* The requirement of Nkx2–1 in the temporal specification of cortical interneuron subtypes. *Neuron* **59,** 722–732 (2008).
36. Taniguchi, H., Lu, J. & Huang, Z. J. The spatial and temporal origin of chandelier cells in mouse neocortex. *Science* **339,** 70–74 (2013). **References 35 and 36 provide support for Nkx2-1 functioning as a master regulator of the fate of specific cortical interneuron identities.**
37. Inan, M., Welagen, J. & Anderson, S. A. Spatial and temporal bias in the mitotic origins of somatostatin- and parvalbumin-expressing interneuron subgroups

and the chandelier subtype in the medial ganglionic eminence. *Cereb. Cortex* **22,** 820–827 (2012).
38. Denaxa, M. *et al*. Maturation-promoting activity of SATB1 in MGE-derived cortical interneurons. *Cell Rep.* **2,** 1351–1362 (2012).
39. Close, J. *et al*. Satb1 is an activity-modulated transcription factor required for the terminal differentiation and connectivity of medial ganglionic eminence-derived cortical interneurons. *J. Neurosci.* **32,** 17690–17705 (2012).
40. Flames, N. *et al*. Delineation of multiple subpallial progenitor domains by the combinatorial expression of transcriptional codes. *J. Neurosci.* **27,** 9682–9695 (2007).
41. Ciceri, G. *et al*. Lineage-specific laminar organization of cortical GABAergic interneurons. *Nature Neurosci.* **16,** 1199–1210 (2013).
42. Brown, K. N. *et al*. Clonal production and organization of inhibitory interneurons in the neocortex. *Science* **334,** 480–486 (2011).
43. Corbin, J. G., Nery, S. & Fishell, G. Telencephalic cells take a tangent: non-radial migration in the mammalian forebrain. *Nature Neurosci.* **4,** 1177–1182 (2001).
44. Marín, O. & Rubenstein, J. L. R. A long, remarkable journey: tangential migration in the telencephalon. *Nature Rev. Neurosci*. **2,** 780–790 (2001).
45. Hobert, O. Specification of the nervous system. In *Wormbook: the Online Review of* C. elegans *Biology* http://www.wormbook.org/ (2005).
46. Beier, K. T., Samson, M. E., Matsuda, T. & Cepko, C. L. Conditional expression of the TVA receptor allows clonal analysis of descendents from Cre-expressing progenitor cells. *Dev. Biol.* **353,** 309–320 (2011).
47. Cancedda, L., Fiumelli, H., Chen, K. & Poo, M. M. Excitatory GABA action is essential for morphological maturation of cortical neurons *in vivo*. *J. Neurosci*. **27,** 5224–5235 (2007).
48. Bortone, D. & Polleux, F. KCC2 expression promotes the termination of cortical interneuron migration in a voltage-sensitive calcium-dependent manner. *Neuron* **62,** 53–71 (2009).
49. De Marco García, N. V., Karayannis, T. & Fishell, G. Neuronal activity is required for the development of specific cortical interneuron subtypes. *Nature* **472,** 351–355 (2011).
References 48 and 49 provide the best evidence so far for a role for activity in the positioning and maturation of cortical interneurons.
50. McKinsey, G. L., Lindtner, S., Trzcinski, B. & Visel, A. Dlx1&2-dependent expression of Zfhx1b (Sip1, Zeb2) regulates the fate switch between cortical and striatal interneurons. *Neuron* **77,** 83–98 (2013).
51. van den Berghe, V. *et al*. Directed migration of cortical interneurons depends on the cell-qutonomous action of Sip1. *Neuron* **77,** 70–82 (2013).
52. Lyons, M. R., Schwarz, C. M. & West, A. E. Members of the myocyte enhancer factor 2 transcription factor family differentially regulate Bdnf transcription in response to neuronal depolarization. *J. Neurosci.* **32,** 12780–12785 (2012).
53. West, A. E. & Greenberg, M. E. Neuronal activity-regulated gene transcription in synapse development and cognitive function. *Cold Spring Harb. Perspect. Biol.* **3,** a005744 (2011).
54. Southwell, D. G., Froemke, R. C., Alvarez-Buylla, A., Stryker, M. P. & Gandhi, S. P. Cortical plasticity induced by inhibitory neuron transplantation. *Science* **327,** 1145–1148 (2010).
55. Bráz, J. M. *et al*. Forebrain GABAergic neuron precursors integrate into adult spinal cord and reduce injury-induced neuropathic pain. *Neuron* **74,** 663–675 (2012).
56. Martínez-Cerdeño, V. *et al*. Embryonic MGE precursor cells grafted into adult rat striatum integrate and ameliorate motor symptoms in 6-OHDA-lesioned rats. *Cell Stem Cell* **6,** 238–250 (2010).
57. Taniguchi, H. *et al*. A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71,** 995–1013 (2011).
58. Hippenmeyer, S. *et al*. A developmental switch in the response of DRG neurons to ETS transcription factor signaling. *PLoS Biol*. **3,** e159 (2005).
59. Douglas, R. J. & Martin, K. A. A functional microcircuit for cat visual cortex. *J. Physiol*. **440,** 735–69 (1991).
60. Douglas, R. J., Koch, C., Mahowald, M. & Martin, K. A. Recurrent excitation in neocortical circuits. *Science* **269,** 981–985 (1995).
61. Buzsáki, G. & Draguhn, A. Neuronal oscillations in cortical networks. *Science* **304,** 1926–1929 (2004).
62. Wang X. J., Tegnér, J., Constantinidis, C., Goldman-Rakic, P. S. Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proc. Natl Acad. Sci. USA.* **101,** 1368–1373 (2004).
63. Silver, R. A. Neuronal arithmetic. *Nature Rev. Neurosci*. **11,** 474–489 (2010).
64. Holt, G. R. & Koch, C. Shunting inhibition does not have a divisive effect on firing rates. *Neural Comput.* **9,** 1001–1013 (1997).
65. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nature Rev. Neurosci.* **13,** 51–62 (2012).
66. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature Neurosci.* **4,** 819–825 (2001).
67. Mitchell, S. J. & Silver, R. A. Shunting inhibition modulates neuronal gain during synaptic excitation. *Neuron* **38,** 433–445 (2003).
68. Tiesinga, P. H. & Sejnowski, T. J. Rapid temporal modulation of synchrony by competition in cortical interneuron networks. *Neural Comput.* **16,** 251–275 (2004).
69. Chance, F. S., Abbott, L. F. & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* **35,** 773–782 (2002).
70. Taniguchi, H., Huang, Z. J. & Callaway, E. M. Contrast dependence and differential contributions from somatostatin- and parvalbumin-expressing neurons to spatial integration in mouse V1. *J. Neurosci*. **33,** 11145–11154 (2013).
71. van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274,** 1724–1726 (1996).
72. Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426,** 442–446 (2003).
73. Haider, B., Duque, A. & Hasenstaub, A. R. Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* **26,** 4535–4545 (2006).
74. Okun, M. & Lampl, I. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature Neurosci.* **11,** 535–537 (2008).
75. Haider, B., Häusser, M. & Carandini, M. Inhibition dominates sensory responses in the awake cortex. *Nature* **493,** 97–100 (2013).
This study demonstrates that during wakefulness inhibition strongly shapes the spatial and temporal response properties of visual cortical neurons.
76. Pouille, F. & Scanziani, M. Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science* **293,** 1159–1163 (2001).
77. Renart, A. *et al*. The asynchronous state in cortical circuits. *Science* **327,** 587–590 (2010).
78. Rudy, B., Fishell, G., Lee, S. & Hjerling-Leffler, J. Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* **71,** 45–61 (2011).
79. Xu, X., Roby, K. D. & Callaway, E. M. Immunochemical characterization of inhibitory mouse cortical neurons: three chemically distinct classes of inhibitory cells. *J. Comp. Neurol.* **518,** 389–340 (2010).
80. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci*. **8,** 1263–1268 (2005).
81. Zhang, F., Aravanis, A. M., Adamantidis, A., de Lecea, L. & Deisseroth, K. Circuit-breakers: optical technologies for probing neural signals and systems. *Nature Rev. Neurosci*. **8,** 577–581 (2007).
82. Somogyi, P., Tamas, G., Lujan, R. & Buhl, E. H. Salient features of synaptic organisation in the cerebral cortex. *Brain Res. Rev.* **26,** 113–135 (1998).
83. Galarreta, M. & Hestrin, S. A network of fast-spiking cells in the neocortex connected by electrical synapses. *Nature* **402,** 72–75 (1999).
84. Kvitsiani, D., Ranade, S., Hangya, B., Taniguchi, H., Huang, J. Z. & Kepecs, A. Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature* **498,** 363–366 (2013).
This study provides evidence that genetically identified interneuron classes are recruited at specific behavioural events.
85. Pfeffer, C. K., Xue, M., He, M., Huang, Z. J. & Scanziani, M. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature Neurosci.* **16,** 1068–1076 (2013).
This paper defines the rules of connectivity for marker-defined interneuron classes.
86. Szabadics, J., Lorincz, A. & Tamás, G. Beta and gamma frequency synchronization by dendritic gabaergic synapses and gap junctions in a network of cortical interneurons. *J. Neurosci.* **21,** 5824–5831 (2001).
87. Royer, S. *et al.* Control of timing, rate and bursts of hippocampal place cells by dendritic and somatic inhibition. *Nature Neurosci*. **15,** 769–775 (2012).
The first direct demonstration of the distinct roles of PV and SST interneurons in awake hippocampus.
88. Lovett-Barron, M. *et al*. Regulation of neuronal input transformations by tunable dendritic inhibition. *Nature Neurosci*. **15,** 423–430 (2012).
89. Losonczy, A., Zemelman, B. V., Vaziri, A. & Magee, J. C. Network mechanisms of theta related neuronal activity in hippocampal CA1 pyramidal neurons. *Nature Neurosci*. **13,** 967–972 (2010).
90. Wilson, N. R., Runyan, C. A., Wang, F. L. & Sur, M. Division and subtraction by distinct cortical inhibitory networks *in vivo*. *Nature* **488,** 343–348 (2012).
91. Atallah, B. V., Bruns, W., Carandini, M. & Scanziani, M. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron* **73,** 159–170 (2012).
92. Lee, S.-H. *et al*. Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature* **488,** 379–383 (2012).
93. Murayama, M., Pérez-Garci, E., Nevian, T., Bock, T., Senn, W & Larkum, M. E. Dendritic encoding of sensory stimuli controlled by deep cortical interneurons. *Nature* **457,** 1137–1141 (2009).
This study reveals how a specific interneuron type gates bursting in layer 5 pyramidal cells.
94. Berger, T. K., Perin, R., Silberberg, G. & Markram, H. Frequency-dependent disynaptic inhibition in the pyramidal network: a ubiquitous pathway in the developing rat neocortex. *J. Physiol.* **587,** 5411–5425 (2009).
95. Kapfer, C., Glickfeld, L. L., Atallah, B. V. & Scanziani, M. Supralinear increase of recurrent inhibition during sparse activity in the somatosensory cortex. *Nature Neurosci*. **10,** 743–753 (2007).
96. Miles, R., Tóth, K., Gulyas, A. I., Hájos, N. & Freund, T. F. Differences between somatic and dendritic inhibition in the hippocampus. *Neuron* **16,** 815–823 (1996).
97. Adesnik, H. & Scanziani, M. Lateral competition for cortical space by layer-specific horizontal circuits. *Nature* **464,** 1155–1160 (2010).
98. Xu, H., Jeong, H. Y., Tremblay, R. & Rudy, B. Neocortical somatostatin-expressing GABAergic interneurons disinhibit the thalamorecipient layer 4. *Neuron* **77,** 155–167 (2013).
99. Hájos, N., Acsády, L. & Freund, T. F. Target selectivity and neurochemical characteristics of VIP-immunoreactive interneurons in the rat dentate gyrus. *Eur. J. Neurosci.* **8,** 1415–1431 (1996).
100. Acsády, L., Görcs, T. J. & Freund, T. F. Different populations of vasoactive intestinal polypeptide-immunoreactive interneurons are specialized to control pyramidal cells or interneurons in the hippocampus. *Neuroscience* **73,** 317–334 (1996).

101. Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G. & Rudy, B. A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nature Neurosci.* **16,** 1662–1670 (2013).
102. Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J. & Kepecs, A. Cortical interneurons that specialize in disinhibitory control. *Nature* **503,** 521–524 (2013).
This study is a direct demonstration that VIP-expressing interneurons are disinhibitory and are recruited by behavioural reinforcers, which together with references 85 and 101 reveals that this function is supported by a microcircuit conserved across regions.
103. Hestrin, S. & Armstrong, W. E. Morphology and physiology of cortical neurons in layer I. *J. Neurosci.* **16,** 5290–5300 (1996).
104. Jiang, X., Wang, G., Lee, A. J., Stornetta, R. L. & Zhu, J. J. The organization of two new cortical interneuronal circuits. *Nature Neurosci.* **16,** 210–218 (2013).
105. Letzkus, J. J. *et al.* A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480,** 331–335 (2011).
This paper demonstrates a functionally relevant disinhibitory circuit in the auditory cortex.
106. Lapray, D. *et al.* Behavior-dependent specialization of identified hippocampal interneurons. *Nature Neurosci.* **15,** 1265–1271 (2012).
107. Varga, C., Golshani, P. & Soltesz, I. Frequency-invariant temporal ordering of interneuronal discharges during hippocampal oscillations in awake mice. *Proc. Natl Acad. Sci. USA* **109,** E2726–E2734 (2012).
This article demonstrates the hippocampal recruitment of distinct interneuron types in awake mice.
108. Gentet, L. J. *et al.* Unique functional properties of somatostatin-expressing GABAergic neurons in mouse barrel cortex. *Nature Neurosci.* **15,** 607–612 (2012).
109. Gentet, L. J., Avermann, M., Matyas, F., Staiger, J. F. & Petersen, C. C. H. Membrane potential dynamics of GABAergic neurons in the barrel cortex of behaving mice. *Neuron* **65,** 422–435 (2010).
110. Csicsvari, J., Hirase, H., Czurko, A. & Buzsáki, G. Reliability and state dependence of pyramidal cell-interneuron synapses in the hippocampus: an ensemble approach in the behaving rat. *Neuron* **21,** 179–189 (1998).
111. Whittington, M. A. & Traub, R. D. Interneuron diversity series: inhibitory interneurons and network oscillations *in vitro*. *Trends Neurosci.* **26,** 676–682 (2003).
112. Buzsáki, G. & Chrobak, J. J. Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Curr. Opin. Neurobiol.* **5,** 504–510 (1995).
113. Klausberger, T., *et al.* Brain-state- and cell-type-specific firing of hippocampal interneurons *in vivo*. *Nature* **421,** 844–848 (2003).
An elegant demonstration of how different interneuron types specialize in specific network oscillations.
114. Klausberger, T., Márton, L. F., Baude, A., Roberts, J., Magill, J. S. & Somogyi, P. Spike timing of dendrite-targeting bistratified cells during hippocampal network oscillations *in vivo*. *Nature Neurosci.* **7,** 41–47 (2004).
115. Klausberger, T. *et al.* Complementary roles of cholecystokinin- and parvalbumin-expressing GABAergic neurons in hippocampal network oscillations. *J. Neurosci.* **25,** 9782–9793 (2005).
116. Tukker, J. J., Fuentealba, P., Hartwich, K., Somogyi, P. & Klausberger, T. Cell type-specific tuning of hippocampal interneuron firing during gamma oscillations *in vivo*. *J. Neurosci.* **27,** 8184–8189 (2007).
117. Klausberger, T. & Somogyi, P. Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. *Science* **321,** 53–57 (2008).
118. Cardin, J. A., *et al.* Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459,** 663–667 (2009).
119. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* **459,** 698–702 (2009).
References 118 and 119 provided the first causal evidence for the role of PV interneurons in gamma oscillations.
120. Mountcastle, V. B., Talbot, W. H., Sakata, H., & Hyvarinen, J. Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys: neuronal periodicity and frequency discrimination. *J. Neurophysiol.* **32,** 452–484 (1969).
121. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* **55,** 131–141 (2007).
122. Isomura, Y., Harukuni, R., Takekawa, T., Aizawa, H. & Fukai, T. Microcircuitry coordination of cortical motor information in self-initiation of voluntary movements. *Nature Neurosci.* **12,** 1586–1593 (2009).
123. Hayden, B. Y., Pearson, J. M. & Platt, M. L. Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neurosci.* **14,** 933–939 (2011).
124. Carandini, M. From circuits to behavior: a bridge too far? *Nature Neurosci.* **15,** 507–509 (2012).
125. Alitto, H. J. & Dan, Y. Cell-type-specific modulation of neocortical activity by basal forebrain input. *Front. Syst. Neurosci.* **6,** 79 (2013).

Acknowledgements Work in the authors' laboratories is supported by grants from the US National Institutes of Health (R01NS075531 to A.K. and MH071679, MH095147, NS074972 and NS081297 to G.F.) and generous support from the McKnight (A.K.) and Simons Foundations (G.F.). We are grateful to G. Buzsaki, C. McBain, B. Rudy, M. Long, R. Tsien and members of our laboratories for discussions and comments. We thank J. Demidschstein for creating Fig. 2. and J. Kuhl for Figs 1 and 3.

Author information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/imruxc. Correspondence should be addressed to G.F. (gordon.fishell@med.nyu.edu) or A.K. (kepecs@cshl.edu).

REVIEW

doi:10.1038/nature12984

The bone marrow niche for haematopoietic stem cells

Sean J. Morrison[1] & David T. Scadden[2]

Niches are local tissue microenvironments that maintain and regulate stem cells. Haematopoiesis provides a model for understanding mammalian stem cells and their niches, but the haematopoietic stem cell (HSC) niche remains incompletely defined and beset by competing models. Recent progress has been made in elucidating the location and cellular components of the HSC niche in the bone marrow. The niche is perivascular, created partly by mesenchymal stromal cells and endothelial cells and often, but not always, located near trabecular bone. Outstanding questions concern the cellular complexity of the niche, the role of the endosteum and functional heterogeneity among perivascular microenvironments.

Haematopoietic stem cell (HSC) niches are present in diverse tissues throughout development, beginning in the aorta–gonad–mesonephros (AGM) region and the yolk sac, followed by the placenta, fetal liver, spleen and bone marrow[1]. Postnatally, the bone marrow is the primary site of HSC maintenance and haematopoiesis, but in response to haematopoietic stress the niche can shift to extramedullary sites. Defining niche components and how they work in concert to regulate haematopoiesis provides the opportunity to improve regeneration following injury or HSC transplantation and to understand how disordered niche function could contribute to disease. In this Review, we focus on the nature of the HSC niche in bone marrow because it is the subject of most of the recent research and controversies.

Historical context

Following Darwin's contributions to evolutionary theory, there was much emphasis on defining hierarchical evolutionary relationships among organisms. Morphological similarities were used to construct ancestral trees that connected complex multicellular organisms to an original monocellular "stem celle"[2]. Lineage relationships were formulated, and biologist Ernst Haeckel proposed that cell organization in a developing organism was the recapitulation of events in the evolution of the species, with cells deriving from a stem cell equivalent[3]. Thirty years later, haematologist Artur Pappenheim proposed a less grand and more accurate formulation based on improved cell-morphology visualization techniques — that cells of the blood were related to one another, with mature cell types descending from a single cell type in a "unified view of haematopoiesis"[4]. In so doing, he articulated the hypothesis of tissue stem cells. This concept took about half a century to define experimentally through the inspired work of James Till and Ernest McCulloch, who showed that single cells could yield multilineage descendants while preserving the multipotency of the mother cell[5–7]. The researchers gave substance to the idea of a stem cell and gave us methods to define the cardinal properties of those cells — self-renewal and differentiation.

Till and McCulloch based much of their work on an *in vivo* spleen colony-forming (CFU-S) assay now known to measure mainly multipotent progenitors rather than long-term self-renewing HSCs[8,9]. The imprecise nature of that assay contributed to the formulation of the niche hypothesis by Ray Schofield in 1978. Recognizing that the putative CFU-S stem cells were less robust than cells of the bone marrow at reconstituting haematopoiesis in irradiated animals, he proposed that a specialized bone marrow niche preserved the reconstituting ability of stem cells[10]. His colleagues at the University of Manchester concurrently sought to define what made bone marrow a nurturing context for HSCs, and haematologist Michael Dexter showed that largely mesenchymal 'stromal' cell cultures could maintain primitive haematopoietic cells *ex vivo*[11]. Furthermore, another colleague, Brian Lord, progressively reamed long bone marrow cavities and showed that primitive cells tended to localize towards the endosteal margins, leading to the hypothesis that bone might regulate haematopoiesis[12] (Fig. 1).

These early studies were followed by *in vitro* evidence that osteoblasts differentiated in culture from human bone marrow stromal cells could produce haematopoietic cytokines and support primitive haematopoietic cells in culture[13]. This fostered the idea that bone cells might create the HSC niche, but it was essential to move to engineered mouse strains to test the hypothesis *in vivo*. Two studies followed, including a mouse model in which a promoter that was restricted in activity to osteoblastic cells was used to drive expression of a constitutively active parathyroid hormone receptor[14]. Along similar lines, Linheng Li's laboratory used a promoter that has since been shown to be restricted in bone marrow stroma to primitive and mature osteolineage cells[15], to delete the *BMPr1a* gene[16]. In both models, the number of endosteal osteoblasts and the number of primitive haematopoietic cells (scored as stem cells given the measures in use at the time) increased. These data provided the first evidence of specific heterologous cells regulating mammalian stem cells *in vivo*, although it remained unclear whether the regulation was direct or indirect. This demonstrated that the niche was experimentally tractable, prompting a series of studies that have since refined our understanding of the complexity of the bone marrow microenvironment.

Studies of the niche have now more precisely determined the components that regulate HSCs, and to some extent other haematopoietic progenitors, in the bone marrow. Like any interactive system there are complex regulatory relationships among cells in the bone marrow. A perturbation in one cell type that leads to an effect in another cell type does not necessarily require the interaction between the cells to be direct. The data now suggest that the early studies that observed effects on HSC frequency as a consequence of genetic manipulation in osteoblastic cells reflected indirect effects rather than the existence of an osteoblastic niche. Indeed, expression of constitutively active parathyroid hormone receptors in osteoblasts[14] probably causes widespread changes in many cell types of the bone marrow, including in the vasculature. Current data suggest

[1]Howard Hughes Medical Institute, Children's Research Institute, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [2]Center for Regenerative Medicine, Massachusetts General Hospital, Harvard Stem Cell Institute and the Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

there are specialized niches for distinct types of haematopoietic stem and progenitor cells, and that each niche may be created by multiple cell types that contribute to the niches in unique as well as redundant ways[17]. Indeed, there is heterogeneity among HSCs themselves[18–20], raising the possibility of cellularly distinct niches for distinct subpopulations of HSCs. This Review focuses on the current data and the unanswered questions.

Mapping the bone marrow space

A niche is defined by anatomy and function[21] — a local tissue microenvironment that directly maintains and regulates a particular kind of stem cell or progenitor[22]. Determining what cells neighbour HSCs and regulate HSC maintenance has been complicated by the difficulty in retaining histological integrity when sectioning bone, as well as the complexity of immunostaining methods that are necessary to identify HSCs.

The identification of markers that reliably identify HSCs *in vivo* was an important step in defining the niche[22]. Despite having the ability to isolate HSCs by flow cytometry for decades[23], identifying HSCs within tissues has remained a challenge because the combination of immunofluorescent markers used to isolate HSCs by flow cytometry was too complex for microscopy. Consequently, markers of poor specificity were often used. For example, putative HSCs have been localized in the bone marrow using retention labels 5-bromodeoxyuridine (BrdU) or GFP-labelled histone H2B (H2B–GFP) as markers[24,25]. Although there is a subset of HSCs that preferentially retains H2B–GFP and BrdU, these markers by themselves have very poor specificity — most bone marrow cells that retain these labels are not HSCs[18,19,26].

When positive staining for CD150 was combined with negative staining for CD48 and CD41, HSCs could finally be highly purified using a simple two-colour stain[27]. All serially transplantable HSCs in young adult mice are contained within the $CD150^{+}CD48^{-}CD41^{-}/CD41^{low}$ population of bone marrow cells, including the most quiescent HSCs[26–29]. This made it possible to localize HSCs in sections through haematopoietic tissues using markers validated to give high purity. Most $CD150^{+}CD48^{-}CD41^{-}$lineage^{-} cells in the bone marrow and spleen localize adjacent to sinusoid vessels, and nearly all are within five cell diameters of a sinusoid[27,30] (Fig. 2). HSCs are five times more likely than other haematopoietic cells to be immediately adjacent to a sinusoid[30]. HSCs are distributed throughout the bone marrow, with less than 20% within 10 μm of the endosteum[27,30–32]. Nonetheless, most HSCs are found in the trabecular region of bone marrow, suggesting that HSCs, or their niche, may be directly or indirectly regulated by factors present near bone surfaces.

The frequent localization of HSCs adjacent to blood vessels suggested that HSCs might be maintained in a perivascular niche by endothelial or perivascular cells[27,33]. But HSCs are mobile, regularly entering and exiting circulation[34]. This raised the possibility that the cells observed near vessels were in transit, perhaps delayed from entering or exiting circulation by

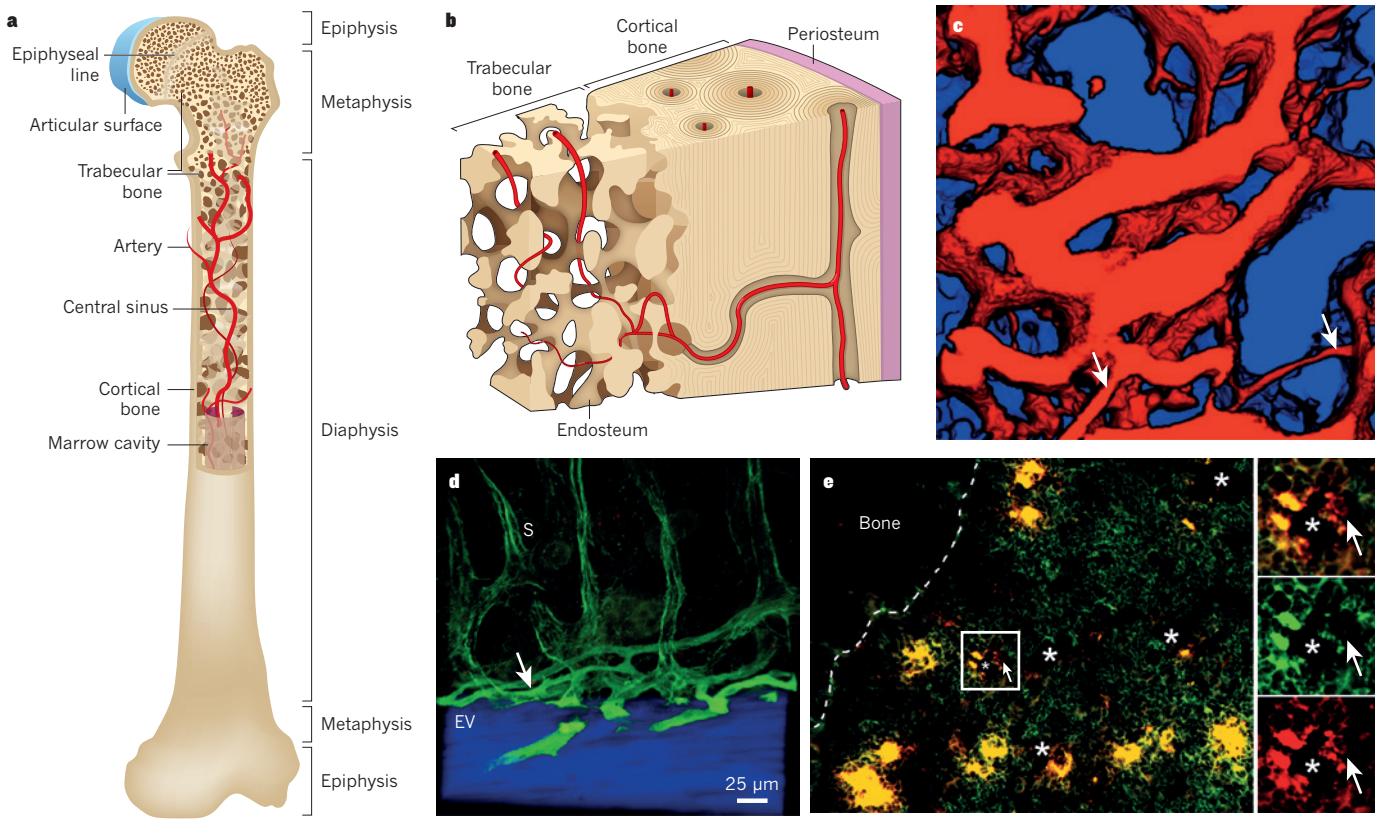


Figure 1 | Bone marrow anatomy. Haematopoietic stem cells (HSCs) reside mainly within bone marrow during adulthood. Bone marrow is a complex organ, containing many different haematopoietic and non-haematopoietic cell types, that is surrounded by a shell of vascularized and innervated bone. **a**, Minute projections of bone (trabeculae) are found throughout the metaphysis such that many cells in this region are close to the bone surface. **b**, The interface of bone and bone marrow is known as the endosteum, which is covered by bone-lining cells that include bone-forming osteoblasts and bone-resorbing osteoclasts. Arteries carry oxygen, nutrients and growth factors into the bone marrow, before feeding into sinusoids, which coalesce as a central sinus to form the venous circulation. Sinusoids are specialized venules that form a reticular network of fenestrated vessels that allow cells to pass in and out of circulation. There is a particularly rich supply of arterioles, as well as sinusoids, near the endosteum. **c**, Three-dimensional reconstructed photomicrograph from the bone marrow towards the endosteal surface (blue) from 50 μm below the surface, revealing the rich network of vessels (red) (image courtesy of C. Lin, J. Spencer and J. Wu). Smaller arteriolar vessels (white arrows) become larger sinusoidal vessels. The field of view is 350 μm × 350 μm. **d**, A cross-sectional view of blood vessels that run along the endosteal surface (EV) and that transition (white arrow) into sinusoids (S) that then course towards the central sinus (adapted with permission from ref. 31). **e**, The bone marrow is cellularly complex with $CD150^{+}CD48^{-}CD41^{-}$lineage^{-} HSCs (arrow) residing in close contact not only with vascular and perivascular cells (*, sinusoid lumens) but also megakaryocytes (large yellow cells) and other haematopoietic cells (image adapted with permission from ref. 125). In the enlargement on the right, CD150 is shown in red and CD48, CD41 and lineage are shown in green.

migrating through vascular barriers. This issue could not be resolved by histological analysis that captures a single moment in time.

The sequential high-resolution imaging of mice assessed the three-dimensional position of cells in the calvarium over time[32,35]. These studies indicated that primitive haematopoietic cells trafficked to specific microdomains of bone marrow blood vessels where the key HSC localization chemokine CXCL12 and the glycoprotein E-selectin were abundant, then remained in these positions for weeks, generating new cells as indicated by the partitioning of a cytosolic dye. When HSCs were visualized after transplantation into irradiated mice they preferentially localized near the endosteum, consistent with that region being particularly relevant for HSC maintenance[32,36]. However, it was subsequently found that irradiation disrupts sinusoids in the bone marrow[37], raising the possibility that the only blood vessels preserved after irradiation are the arteriolar vessels near the endosteum. Therefore, the peri-endosteal localization of HSCs in these experiments may have reflected, in part, the destruction of sinusoidal niches by irradiation. Overall, the localization data emphasized the possibility of a perivascular niche. How could this be resolved with historical data suggesting that the endosteum and osteoblasts were niche participants?

Osteoblasts are more harbinger than host

Although osteoblastic cells were the first cell population shown to influence haematopoietic stem or progenitor cell frequency when perturbed *in vivo*[14,16], several lines of evidence raised concerns that the effect may not be direct. First, *in vivo* imaging studies using validated markers or labelled stem cells found few HSCs in contact with osteoblastic cells[27,31,32,33,38]. Second, studies that depleted osteoblasts by *Bgn* deficiency[30] or osteoblastic cells by treatment with ganciclovir[39,40] or that increased osteoblasts by strontium treatment[41] had no acute effect on HSC frequency. The studies in which osteoblastic cells were conditionally deleted by ganciclovir showed acute depletion of B lymphoid progenitors and only later showed a decline in a stem/progenitor cell population[39,40]. Third, genetic modification of primitive osteolineage cells had an effect on HSC proliferation and differentiation, but the same modification in mature osteoblasts did not[42]. Finally, a key adhesion molecule thought to mediate osteoblast–HSC interaction, N-cadherin, was called into question.

N-cadherin$^+$ HSCs were proposed to adhere to N-cadherin$^+$ osteoblasts by homophilic adhesion[16,36], promoting HSC maintenance[43–45]; however, these studies did not test whether deletion of the gene that encodes N-cadherin (*Cdh2*) affected HSC function. The levels of N-cadherin staining in HSCs were difficult to distinguish from background fluorescence and depended on anti-N-cadherin antibodies that gave nonspecific staining in some haematopoietic cells[46]. Other studies failed to detect N-cadherin expression by HSCs using gene expression profiling[27,47,48] (https://gexc.stanford.edu/model/3/gene/Cdh2) quantitative reverse-transcription–PCR, flow cytometry with multiple anti-N-cadherin antibodies, western blot, or *Cdh2*:*LacZ* genetrap mice[18,28,30,38]. Conditional deletion of *Cdh2* from HSCs or from osteoblast lineage cells had no effect on HSC frequency, HSC function or haematopoiesis[38,49,50]. Collectively, these data undermined the notion of an N-cadherin$^+$ 'osteoblastic' niche.

Is there any role for osteoblasts or osteolineage cells in HSC regulation? Several lines of evidence suggest that this possibility remains viable but not as it was initially foreseen. First, higher numbers of HSCs reside in the trabecular rich metaphysis[31,51]. This may simply reflect other components of bone marrow co-localizing with bony surfaces; however, conditional deletion of *Sp7* (*Osterix*) results in chondrocytes without osteoblastic differentiation, increasing blood vessels and mesenchymal progenitors in the

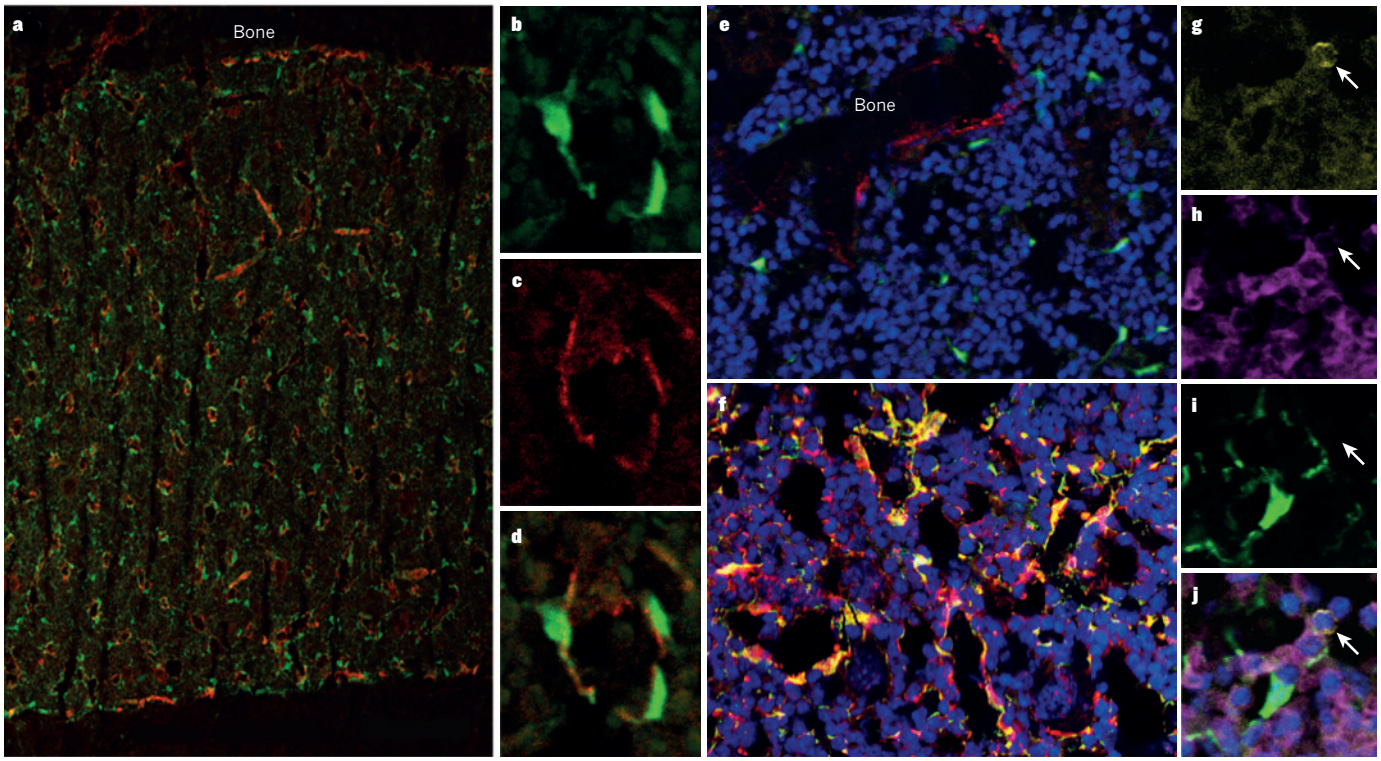


Figure 2 | Haematopoietic stem cells and their niche cells surround sinusoids throughout the bone marrow. **a**, Sections through the bone marrow of mice in which GFP had been knocked in to the *Scf* locus show that Scf-expressing cells (green) include mesenchymal stromal cells and endothelial cells that surround sinusoids and potentially other blood vessels throughout the bone marrow[64]. High magnification shows that **b**, *Scf*-GFP expression overlaps with **c**, the endothelial marker endoglin (shown by **d**, a merge of the two) but also extends beyond the endoglin on the abluminal side of the sinusoids, indicating expression by mesenchymal stromal cells. **e**, *Scf*-GFP (green) is not expressed by osteopontin$^+$ bone-lining cells (red) around trabecular bone, but it is expressed by some nearby perivascular cells. **f**, *Cxcl12*-DsRed (red) exhibits a similar expression pattern, primarily by perivascular mesenchymal cells and endothelial cells around sinusoids throughout the bone marrow, in a pattern that strongly overlaps with *Scf*-GFP (green) in mice with *DsRed* knocked into the *Cxcl12* locus and GFP knocked into the *Scf* locus[17]. Cells that are (**g**) CD150$^+$ and (**h**) CD48 and lineage$^-$ are usually found immediately adjacent to (**i**) *Scf*-GFP$^+$ perivascular cells in the bone marrow. **j**, A merge of **g**–**i**. Images adapted with permission from refs 17 and 64.

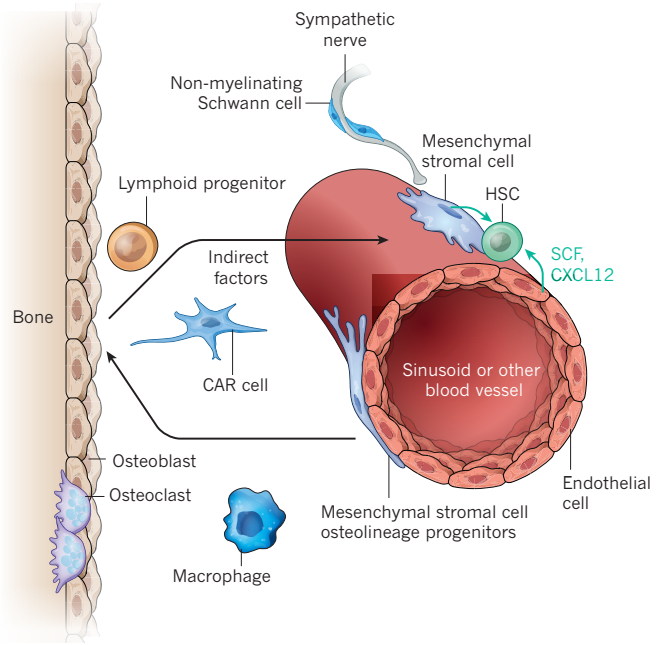


Figure 3 | Haematopoietic stem cells (HSCs) and restricted haematopoietic progenitors occupy distinct niches in the bone marrow. HSCs are found mainly adjacent to sinusoids throughout the bone marrow[27,30,31,33], where endothelial cells and mesenchymal stromal cells promote HSC maintenance by producing SCF[64], CXCL12 (refs 17, 33, 62) and probably other factors. Similar cells may also promote HSC maintenance around other types of blood vessels, such as arterioles. The mesenchymal stromal cells can be identified based on their expression of *Lepr*-Cre[64], *Prx1*-Cre[62], *Cxcl12*-GFP[33] or *Nes*-GFP transgenes[63] in mice and similar cells are likely to be identified by CD146 expression in humans[54]. Perivascular stromal cells, which probably include Cxcl12-abundant reticular (CAR) cells[33], are fated to form bone *in vivo*, express Mx-1-Cre and overlap with CD45/Ter119$^-$PDGFRα$^+$Sca-1$^+$ stromal cells that are highly enriched for mesenchymal stromal cells in culture[66]. It is likely that other cells also contribute to this niche, these probably include cells near bone surfaces in trabecular-rich areas. Other cell types that regulate HSC niches include sympathetic nerves[91,92], non-myelinating Schwann cells (which are also Nes$^+$)[96], macrophages[95] and osteoclasts[97]. The extracellular matrix[120,121] and calcium[56] also regulate HSCs. Osteoblasts do not directly promote HSC maintenance but do promote the maintenance and perhaps the differentiation of certain lymphoid progenitors by secreting CXCL12 and probably other factors[13,17,39,40]. Early lymphoid restricted progenitors thus reside in an endosteal niche that is spatially and cellularly distinct from HSCs.

bone marrow but almost eliminating haematopoiesis in the metaphysis[52]. These data argue that the presence of mature or maturing osteolineage cells in regions with abundant endosteum is crucial for haematopoiesis. Indeed, mesenchymal progenitors that are capable of forming bone are sufficient to create bony ossicles that become invested by host vasculature and HSCs[53,54]. This suggests that bone or bone-forming progenitors can promote the formation or maintenance of HSC niches (for example by recruiting vasculature to the bone marrow) even if they do not directly promote HSC maintenance.

Transplanted haematopoietic stem or progenitor cells preferentially localize to blood vessels in endosteal regions even without prior cytotoxic conditioning[55]. Within the endosteal region, transplanted HSCs position themselves closer to the endosteal surface than progenitor cells[32]. These may again reflect indirect effects of bone-forming osteolineage cells as bone turnover results in high local concentrations of ionic calcium, and the calcium-sensing receptor promotes bone marrow engraftment by HSCs during development or after transplantation[56]. Osteolineage cells also elaborate cytokines and extracellular matrix proteins that may influence a wide range of cell types, some of which may directly regulate HSC function. This is exemplified by parathyroid hormone receptor activation, which induces expression of multiple regulatory molecules (such as IL-6, RANKL and Jagged1) by osteoblasts that can influence other cells in the bone marrow, including the vasculature[14,57]. In addition, osteoblastic expression of transgenes encoding the Wnt antagonists Dkk1 and Wif1 depletes HSCs[58,59]. Finally, the depletion of osteocalcin-expressing cells (osteoblasts or osteocytes) *in vivo* results in an inability to mobilize at least short-term repopulating cells to the blood using granulocyte colony-stimulating factor[60,61] despite osteoblasts having little expression of CXCL12 (refs 17, 33, 62). In aggregate, these data indicate that the endosteal region is important for haematopoiesis, but the mature osteolineage cells probably have an indirect role in modulating HSCs. However, these cells seem to be more important in directly regulating restricted progenitors, a topic discussed later. It is important, therefore, to refocus attention on the endosteum as a regulatory region and not on the osteoblasts themselves (Fig. 3).

The endosteum has a diverse group of cells and anatomical elements, including a rich endowment of arteriolar and sinusoidal blood vessels[31,32] (Fig. 1). The cells include endothelial cells as well as mesenchymal cells with osteolineage potential. These mesenchymal cells reside perivascularly but traffic to the endosteal surface to differentiate into osteoblasts. Undifferentiated mesenchymal cells around blood vessels may promote HSC maintenance throughout the bone marrow, but the mesenchymal cells around vessels in the endosteal region may differ from those distant from endosteal surfaces.

Perivascular regulators of HSCs

Given the localization of HSCs near blood vessels, it was crucial to define the stromal cells surrounding the vessels and to test whether they promote HSC maintenance. Attention focused on the mesenchymal cells that surround blood vessels throughout the bone marrow. Although mesenchymal stroma are likely to be heterogeneous, and the precise relationships between cells expressing various markers remain to be defined, perivascular mesenchymal cells that express CD146 in humans[54] and *Cxcl12*-GFP[33], *Nes*-GFP[63], full length *Lepr*-Cre[64], *Prx-1*-Cre[62], *Sp7*-Cre[62] and inducible *Mx-1*-Cre[15] mice all generate osteoblastic cells and all express factors that promote HSC maintenance. CXCL12-abundant 'reticular' (CAR) cells adjacent to sinusoids were first shown to co-localize with HSCs throughout the bone marrow[33]. Ablation of Cxcl12-expressing bone marrow cells depletes HSCs as well as severely impairing the adipogenic and osteogenic capacity of bone marrow cells[65]. Human CD146$^+$ skeletal stem cells also localize adjacent to sinusoids in the bone marrow and synthesize high levels of the HSC niche factors stem cell factor (SCF) and CXCL12 (ref. 54).

The possibility that mesenchymal stem or stromal cells (MSCs) are part of the HSC niche was further supported by the finding that MSCs in the bone marrow express a *Nes*-GFP transgene and localize around blood vessels throughout the bone marrow[63]. HSCs commonly localize adjacent to *Nes*-GFP$^+$ cells and these cells express high levels of Scf and Cxcl12. Moreover, fibroblast activation protein (FAP) is expressed by bone marrow stromal cells that have many characteristics of MSCs, including Cxcl12, Scf, Pdgfrα and Sca-1 expression[66,67], and ablation of these FAP$^+$ cells leads to bone marrow hypocellularity, anaemia and depletion of osteogenic cells[68,69]. These studies provided strong evidence that MSCs are one component of a perivascular niche for HSCs.

Endothelial cells also contribute to the perivascular HSC niche[27]. The earliest functional evidence supporting this possibility was the observation that conditional deletion of the gene that encodes the gp130 cytokine receptor in endothelial cells led to bone marrow hypocellularity and a reduction in HSC numbers[70]. Inhibition of VEGFR2 signalling in irradiated mice using a blocking antibody impaired the regeneration of sinusoidal endothelial cells and prevented the recovery of LSK stem or progenitor cells as well as CFU-S cells[37]. Endothelial cells can promote HSC maintenance in culture[71], and bone marrow sinusoidal endothelial cells promote long-term reconstituting HSC expansion in culture[72,73]. It has been suggested that E-selectin is exclusively expressed by endothelial cells in the bone marrow, and deficiency of the gene that encodes it renders HSCs more quiescent and resistant to irradiation[74]. These studies suggested that endothelial cells are one component of the HSC niche, but did not address whether they directly or indirectly regulate HSC maintenance *in vivo*.

To formally identify the niche cells, researchers examined which cell populations were the key sources of factors that promote HSC maintenance *in vivo*. For example, SCF has a non-cell-autonomous role for HSC maintenance *in vivo*[75–79]. Differential splicing and proteolytic cleavage yield membrane-bound and soluble forms of SCF. HSCs are depleted in Sl/Sl^{d} mutant mice[80], which express soluble SCF but not the membrane-bound form, indicating that this form is necessary for HSC maintenance[81]. Importantly, mice with a mixture of wild-type and Sl/Sl^{d} stromal cells only exhibit normal haematopoiesis in the immediate vicinity of the wild-type cells, demonstrating that SCF acts locally in creating the niche[82]. Because cell–cell contact is needed between HSCs and those that synthesize SCF, the niche could be localized by identifying the key sources of SCF for HSC maintenance.

Analysis of the Scf expression pattern in mice in which GFP is knocked in to the *Scf* locus revealed that Scf is expressed perivascularly, mainly around sinusoids throughout the bone marrow[64]. Lepr^{+} perivascular stromal cells expressed the highest levels of Scf and endothelial cells expressed lower levels. Gene expression profiling suggested that these Lepr-expressing perivascular cells were mesenchymal. Scf-GFP expression could not be detected in osteoblasts or in haematopoietic cells. Conditional deletion of Scf from perivascular stromal cells (*Lepr*-Cre) or endothelial cells (*Tie2*-Cre) depleted HSCs[64]. However, deletion of Scf from haematopoietic cells (*Vav1*-Cre), osteoblastic cells (*Col2.3*-Cre) and Nestin-expressing perivascular stromal cells (*Nes*-Cre and *Nes*-CreER) did not affect HSC frequency[64]. These results proved there is a perivascular niche for HSCs in which endothelial cells and mesenchymal cells promote HSC maintenance by synthesizing SCF (Fig. 3).

It has been proposed that the endosteal region and its osteoblastic cells provide a unique zone for the maintenance of quiescent HSCs. However, when *Scf* was conditionally deleted from both endothelial cells and perivascular mesenchymal cells in *Lepr-cre*, *Tie2-cre*, $Scf^{fl/-}$ mice, 85% of all long-term multilineage reconstituting cells, including all serially transplantable HSCs and all HSCs in the most quiescent subpopulation, were eliminated[20]. Therefore, even the most primitive and quiescent HSCs are maintained by a perivascular niche. Whether there are functionally distinct perivascular niches in different regions of the bone marrow, such as in the endosteal region, remains an open question.

Are other key niche factors also synthesized primarily by perivascular cells? Cxcl12 is a chemokine that is required for HSC maintenance and HSC retention in the bone marrow[33,83–86]. Global deletion of *Cxcl12*, or the gene that encodes the Cxcl12 receptor, *Cxcr4*, depletes HSCs from the bone marrow[33,83,87]. Cxcl12 is primarily expressed by perivascular mesenchymal stromal cells (CAR cells, *Nes*-GFP, *Lepr*-Cre or *Prx1*-Cre expressing cells), with 100-fold lower levels of expression in endothelial cells and 1,000-fold lower levels in osteoblasts[17,33,62,88,89]. Conditional deletion of *Cxcl12* from perivascular mesenchymal cells using *Prx1*-Cre and *Lepr*-Cre depleted and mobilized HSCs, respectively[17,62]. HSCs were depleted but not mobilized when *Cxcl12* was conditionally deleted from endothelial cells (*Tie2*-Cre)[17,62]. HSC frequency and bone marrow retention were not affected when *Cxcl12* was conditionally deleted from osteoblasts or their progenitors (*Col2.3*-Cre and *Sp7*-Cre), haematopoietic cells (*Vav1*-Cre), or *Nes*-Cre-expressing stromal cells[17,62]. These data confirmed that HSCs reside in a perivascular niche in which mesenchymal stromal cells and endothelial cells each synthesize multiple factors that promote HSC maintenance and localization.

Although conditional deletion of *Scf* and *Cxcl12* with *Nes*-Cre and *Nes*-CreER did not have any effect on HSC frequency[17,64], *Nes*-GFP^{+} perivascular cells are almost certainly part of the HSC niche[63]. Each of these *Nes* alleles are transgenes with different expression patterns in the bone marrow[64]. *Nes*-Cre seems not to be expressed in the bone marrow and *Nes*-CreER exhibits very limited perivascular expression that does not resemble the expression patterns of *Scf*-GFP, *Cxcl12*-DsRed, *Nes*-GFP or *Nes*-Cherry[64]. However, *Nes*-GFP expression strongly overlaps with *Lepr*-Cre expression by perivascular cells throughout the bone marrow[64,67,90]. Thus, it is likely that *Nes*-GFP^{+} perivascular MSCs are a component of the HSC niche even though *Nes*-Cre mediated deletion of *Scf* or *Cxcl12* did not deplete HSCs. Going forward, it will be useful to identify other Cre alleles that are specifically expressed in *Nes*-GFP^{+} cells to compare their function with other perivascular stromal cells.

Complexity of the perivascular HSC niche

Endothelial cells and mesenchymal stromal cells are not the only cell types that regulate the perivascular HSC niche (Fig. 3). The sympathetic nervous system regulates CXCL12 expression and HSC retention in the bone marrow[91,92]. This seems to be accomplished by sympathetic nerve fibres that likely synapse on perivascular cells around a subset of blood vessels in the bone marrow, conferring circadian regulation of CXCL12 expression and HSC mobilization. Circadian oscillation in the clearance of aged neutrophils by macrophages in the bone marrow also contributes to these circadian changes in CXCL12 expression and HSC circulation[93]. Consistent with this, macrophages modulate Cxcl12 expression by *Nes*-GFP^{+} cells and HSC retention in the bone marrow[94,95]. Non-myelinating Schwann cells seem to regulate the niche by regulating TGF-β activation and potentially by secreting other factors[96]. Osteoclasts, or osteoclast activity at the endosteum, may also influence HSC maintenance and bone marrow retention[56,97,98]. Many different cell types are likely to directly or indirectly regulate the perivascular HSC niche.

Given the complexity of cell types implicated in the regulation of HSCs, there is no singular niche cell. Rather, the niche integrates the function of multiple participants. It is important to bear in mind that niche composition and niche function may change under different physiological conditions or in response to stress. It is also important to note that many of the Cre recombinase alleles used so far to study niche cells were active during development. Although this was necessary to achieve efficient gene deletion (temporally regulated CreER alleles tend to give much lower levels of recombination) and no abnormalities in development were noted, indirect effects on surrounding cell types and compensatory changes cannot be excluded. Even though endothelial cells and perivascular mesenchymal cells express SCF and CXCL12, conditional deletion of these factors from these cell types may have direct and indirect effects on HSCs.

There may also be long-range signals circulating through the blood that regulate HSC or niche function, perhaps integrating stem-cell activity with overall physiology[99]. These may include hormones that signal reproductive or nutritional status, or even haematopoietic cytokines. For example, thrombopoietin is required for HSC maintenance[100–103]. The main sites of thrombopoietin synthesis are in the liver and kidney, although it is also synthesized at lower levels by bone marrow stroma[104,105]. Conditional deletion experiments will be required to determine the physiologically important source or sources of thrombopoietin for HSC maintenance.

There may also be functionally distinct perivascular environments in the bone marrow based on vessel type. Most studies of perivascular niches in the bone marrow have focused on sinusoids because they are the most abundant blood vessels in the bone marrow and most HSCs, MSCs, Scf-expressing cells and Cxcl12-expressing cells are in close proximity to them[17,27,33,54,64]. However, other types of blood vessel, such as arterioles, may have an important role in HSC maintenance. A recent study reported that NG2^{+}, but Lepr^{-}, mesenchymal cells that surround arterioles in the bone marrow are important for the maintenance of quiescent HSCs[106]. This conclusion was based on the observation that HSCs were depleted and driven into cycle when *NG2*-CreER^{+} cells were ablated by treatment with diphtheria toxin. However, these data seem to conflict with the earlier observation that quiescent HSCs are eliminated from the bone marrow when *Scf* is conditionally deleted using *Tie2*-Cre and *Lepr*-Cre, which recombine in endothelial cells and mesenchymal cells that are primarily around sinusoids throughout the bone marrow[20]. It will thus be interesting to determine whether HSCs are depleted when *Scf* is conditionally deleted using *NG2*-CreER or whether *NG2*-CreER is expressed by cells other than periarteriolar cells in the bone marrow. Similarly, it will be important to assess whether *Lepr*-expressing perivascular cells contribute to arteriolar niches in the bone marrow. In the end, perivascular niches associated with both sinusoids and arterioles may regulate HSC maintenance and quiescence in the bone marrow. Dissecting the diversity in perivascular

environments will require Cre alleles that are specifically expressed within distinct perivascular domains to map their functions.

Evidence of HSCs residing within relatively hypoxic domains within the bone marrow[31,107] has partly been based on staining with pimonidazole[31]. Stained HSCs often reside adjacent to sinusoids in the bone marrow and are found next to cells that do not stain with pimonidazole[31]. This suggests that pimonidazole staining does not reflect ambient oxygen or that it is cell-autonomously determined, rather than reflecting a hypoxic environment. Pimonidazole responds to reducing intermediates and may reflect more about the metabolic state of cells than ambient oxygen levels.

Some researchers have interpreted the dependence of HSC maintenance on HIF-1α as suggestive of a hypoxic niche[108]. However, a number of factors other than hypoxia regulate HIF-1α expression. A recent imaging study using a nanoprobe specifically reflective of ambient oxygen found that oxygen tension was lowest around sinusoids and highest near the endosteum[109]. The entire bone marrow space had much lower levels of oxygen compared with vessels entering the bone marrow, a feature that was largely lost when haematopoiesis was ablated by cytotoxic drugs. It is therefore likely that consumption of oxygen during haematopoiesis renders the marrow hypoxic but that no distinct hypoxic region exists at the endosteum.

Distinct haematopoietic progenitors have distinct niches

HSCs reside within a specialized niche that is distinct from the niches that nurture other haematopoietic progenitors. For example, although osteolineage cells do not directly regulate HSC maintenance, they do regulate some B-cell lineage progenitors. Cultures enriched for osteoblasts support B lymphopoiesis and ablation of osteoblastic cells in adult mice acutely depletes some B lymphoid progenitors[40,110]. Deletion of $G_s\alpha$ in osteoblastic cells, which is necessary for parathyroid hormone receptor signalling, markedly depleted pro- and pre-B cells in a way that could be mitigated with IL-7 (ref. 111). About 30% of $IL7R^+lineage^-$ bone marrow cells, which are enriched for early lymphoid progenitors, localize immediately adjacent to bone-lining cells at the endosteum[17]. Conditional deletion of CXCL12, a factor that promotes the proliferation and maintenance of B-lineage progenitors[84,112] and common lymphoid progenitors (CLPs)[113], in osteoblastic cells depleted CLPs and certain other early lymphoid progenitors from the bone marrow without any effect on HSCs[17]. Therefore, some early lymphoid progenitors depend on an osteoblastic niche that is cellularly and functionally distinct from the perivascular niche that maintains HSCs.

Other lineage-restricted niches may also exist. For example, macrophages seem to be crucial for erythroid maturation, and macrophage depletion reduces normal and malignant erythropoiesis[114]. Other cellular components of the erythropoiesis niche will have to be identified to understand the relationship between this niche and HSC and lymphoid progenitor niches.

The approach of conditionally deleting specific niche factors from candidate niche cells and then examining the consequences for stem or progenitor cell maintenance *in vivo* offers the opportunity to map the niches for each stem cell and restricted progenitor in the haematopoietic system, limited only by the precision of the Cre alleles that are available.

Novel niche factors

In contrast to stem cells in some tissues, HSCs cannot be sustainably expanded in culture. This has impeded our ability to safely and effectively transplant HSCs in certain clinical contexts, such as during gene therapy, in which it would be useful to expand transfected HSCs in culture and verify the quality of the transfected HSCs before transplantation. One possibility why HSCs cannot be expanded in culture is the existence of, so far, unidentified growth factors that are synthesized by the niche *in vivo*.

Some HSC niche factors have only recently been discovered. The addition of pleiotrophin to culture promotes HSC maintenance[115] and *Ptn* deficiency is associated with HSC depletion and impaired haematopoietic regeneration after myelosuppression[116]. Pleiotrophin is synthesized by sinusoidal endothelial cells and *Cxcl12*-expressing perivascular stromal cells, and has a non-cell-autonomous role in promoting HSC function[116]. The Slit receptor Robo4, which is expressed by HSCs and endothelial cells, regulates HSC localization in the bone marrow[117,118]. The *Slit2* ligand is restricted to MSCs and possibly other osteoblast lineage cells. This suggests that pleiotrophin and Robo4–Slit2 are important elements of the perivascular niche. The glycoprotein tenascin-C[119], osteopontin[120,121] and non-canonical Wnts[25] have also been reported to positively or negatively influence HSC numbers in the bone marrow and are among a number of factors that bear further characterization in terms of cellular source or role with respect to the niche.

Perspective

Ten years of experimentation has validated the niche concept and answered some first order questions about the molecular and cellular nature of the HSC niche in the bone marrow. The 'parts' list that make up this niche remains incomplete, but with the pace of current work it is likely that additional components will be defined and ambiguity about overlapping cell populations resolved over the next few years. This will make it possible to compare anatomically and developmentally distinct HSC niches that have different functions. The number of HSCs expands daily within the fetal liver but is sustained at nearly constant levels in the bone marrow, at least in the absence of injury. How components of these niches compare may inform methods for achieving HSC expansion. Similarly, comparing homologous niches among species, such as long-lived humans with short-lived mice, may provide insight into mechanisms for preserving the integrity of haematopoiesis under stress or in response to ageing. Finally, comparing niches among tissues will assess whether the mesenchymal and endothelial populations in brain, gut and skin share characteristics and functions with those defined in the bone marrow. Do diverse adult tissues consistently have perivascular niches for stem-cell maintenance? Do regenerative tissues have niches with common mechanisms for preserving self-renewal? Are there common components that can be engineered into niches *ex vivo*?

With the detail now emerging in our understanding of the bone marrow niche, a number of second order questions can be addressed. Increasingly, niche cells can be genetically tagged or modified, allowing both quantification and molecular manipulation. Coupled with high resolution real-time imaging and well-validated methods to measure haematopoiesis, it is becoming possible to systematically elucidate how the niche responds to stresses or physiological changes to mediate changes at the stem-cell and tissue levels. When stressed by infection, myeloablation or neoplasia, what niche components change in number or function to modify haematopoiesis? Is there a hierarchy of niche components that determine these responses? Can such information allow predictive algorithms that guide specific interventions to achieve desired outcomes?

Another set of questions concerns the manner in which the niche participates in diseases of stem-cell failure, such as aplastic anaemia or neoplasia. The niche may be hostile to normal progenitors in those disease states and, with neoplasia, undergo a facultative response to support altered haematopoiesis[122]. Can changes in the niche be a primary but non-cell-autonomous driver of neoplasia in humans as has been suggested by animal models[42,123,124]? The potential for unravelling how the microenvironment participates in normal and disease physiology is at hand and promises new approaches to haematological disorders. ■

Received 16 August; accepted 5 November 2013.

1. Mikkola, H. K. & Orkin, S. H. The journey of developing hematopoietic stem cells. *Development* **133,** 3733–3744 (2006).
2. Haeckel, E. H. P. A. *Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Decendenz-Theorie.* (1866).
3. Haeckel, E. *The Riddle of the Universe* (Die Weltraetsel, 1895–1899). 1992 Reprint Edition edn, (Prometheus, 1901).
4. Pappenheim, A. Ueber Entwickelung und Ausbildung der Erythroblasten. *Virchows Arch.* **145,** 587–643 (1896).
5. Till, J. E. & McCulloch, E. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14,** 213–222 (1961).
6. Siminovitch, L., McCulloch, E. A. & Till, J. E. The distribution of colony-forming

cells among spleen colonies. *J. Cell. Physiol.* **62,** 327–336 (1963).
7. Siminovitch, L., Till, J. E. & McCulloch, E. A. Decline in colony-forming ability of marrow cells subjected to serial transplantation into irradiated mice. *J. Cell. Physiol.* **64,** 23–31 (1964).
8. Jones, R. J. *et al.* Characterization of mouse lymphohematopoietic stem cells lacking spleen colony-forming activity. *Blood* **88,** 487–491 (1996).
9. Spangrude, G. J., Brooks, D. M. & Tumas, D. B. Long-term repopulation of irradiated mice with limiting numbers of purified hematopoietic stem cells: *in vivo* expansion of stem cell phenotype but not function. *Blood* **85,** 1006–1016 (1995).
10. Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* **4,** 7–25 (1978).
This article describes the niche hypothesis.
11. Dexter, T. M., Allen, T. D. & Lajha, L. G. Conditions controlling the proliferation of hemopoietic stem cells *in vitro*. *J. Cell. Physiol.* **91,** 335–344 (1977).
12. Lord, B. I., Testa, N. G. & Hendry, J. H. The relative spatial distributions of CFUs and CFUc in the normal mouse femur. *Blood* **46,** 65–72 (1975).
13. Taichman, R. S. & Emerson, S. G. Human osteoblasts support hematopoiesis through the production of granulocyte colony-stimulating factor. *J. Exp. Med.* **179,** 1677–1682 (1994).
14. Calvi, L. M. *et al.* Osteoblastic cells regulate the haematopoietic stem cell niche. *Nature* **425,** 841–846 (2003).
In this paper, researchers identified heterologous cells influencing stem/progenitor cells in mammals, providing experimental evidence for the niche hypothesis.
15. Park, D. *et al.* Endogenous bone marrow MSCs are dynamic, fate-restricted participants in bone maintenance and regeneration. *Cell Stem Cell* **10,** 259–272 (2012).
16. Zhang, J. *et al.* Identification of the haematopoietic stem cell niche and control of the niche size. *Nature* **425,** 836–841 (2003).
This study identified heterologous cells that influence stem or progenitor cells in mammals, providing experimental evidence for the niche hypothesis.
17. Ding, L. & Morrison, S. J. Haematopoietic stem cells and early lymphoid progenitors occupy distinct bone marrow niches. *Nature* **495,** 231–235 (2013).
Systematic analysis of CXCL12-expressing cells in the bone marrow demonstrating that stem cells and restricted progenitors depend on cellularly distinct niches.
18. Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nature Biotechnol.* **27,** 84–90 (2009).
19. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135,** 1118–1129 (2008).
20. Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13,** 102–116 (2013).
21. Scadden, D. T. The stem-cell niche as an entity of action. *Nature* **441,** 1075–1079 (2006).
22. Morrison, S. J. & Spradling, A. C. Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. *Cell* **132,** 598–611 (2008).
23. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241,** 58–62 (1988).
24. Arai, F. *et al.* Tie2/angiopoietin-1 signaling regulates hematopoietic stem cell quiescence in the bone marrow niche. *Cell* **118,** 149–161 (2004).
25. Sugimura, R. *et al.* Noncanonical wnt signaling maintains hematopoietic stem cells in the niche. *Cell* **150,** 351–365 (2012).
26. Kiel, M. J. *et al.* Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU. *Nature* **449,** 238–242 (2007).
27. Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121,** 1109–1121 (2005).
This article provides a definition of an immunophenotypic signature for highly enriched stem cells that permitted histological mapping of HSCs within the bone marrow and suggested the existence of a perivascular niche.
28. Morita, Y., Ema, H. & Nakauchi, H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med.* **207,** 1173–1182 (2010).
29. Yilmaz, O. H., Kiel, M. J. & Morrison, S. J. SLAM family markers are conserved among hematopoietic stem cells from old and reconstituted mice and markedly increase their purity. *Blood* **107,** 924–930 (2006).
30. Kiel, M. J., Radice, G. L. & Morrison, S. J. Lack of evidence that hematopoietic stem cells depend on N-cadherin-mediated adhesion to osteoblasts for their maintenance. *Cell Stem Cell* **1,** 204–217 (2007).
31. Nombela-Arrieta, C. *et al.* Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nature Cell Biol.* **15,** 533–543 (2013).
32. Lo Celso, C. *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* **457,** 92–96 (2009).
33. Sugiyama, T., Kohara, H., Noda, M. & Nagasawa, T. Maintenance of the hematopoietic stem cell pool by CXCL12–CXCR4 chemokine signaling in bone marrow stromal cell niches. *Immunity* **25,** 977–988 (2006).
This study identified a perivascular stromal cell that promoted HSC maintenance.
34. Wright, D. E., Wagers, A. J., Gulati, A. P., Johnson, F. L. & Weissman, I. L. Physiological migration of hematopoietic stem and progenitor cells. *Science* **294,** 1933–1936 (2001).
35. Sipkins, D. A. *et al.* *In vivo* imaging of specialized bone marrow endothelial microdomains for tumour engraftment. *Nature* **435,** 969–973 (2005).
This paper reports evidence that subregions of the microvasculature express high levels of Cxcl12 where transplanted haematopoietic progenitors localize and increase in number.
36. Xie, Y. *et al.* Detection of functional haematopoietic stem cell niche using real-time imaging. *Nature* **457,** 97–101 (2009).
37. Hooper, A. T. *et al.* Engraftment and reconstitution of hematopoiesis is dependent on VEGFR2-mediated regeneration of sinusoidal endothelial cells. *Cell Stem Cell* **4,** 263–274 (2009).
The authors of this study found that sinusoidal endothelial cells have specialized features and are necessary for HSC engraftment.
38. Kiel, M. J., Acar, M., Radice, G. L. & Morrison, S. J. Hematopoietic stem cells do not depend on N-cadherin to regulate their maintenance. *Cell Stem Cell* **4,** 170–179 (2009).
39. Visnjic, D. *et al.* Hematopoiesis is severely altered in mice with an induced osteoblast deficiency. *Blood* **103,** 3258–3264 (2004).
40. Zhu, J. *et al.* Osteoblasts support B lymphocyte commitment and differentiation from hematopoietic stem cells. *Blood* **109,** 3706–3712 (2007).
41. Lymperi, S. *et al.* Strontium can increase some osteoblasts without increasing hematopoietic stem cells. *Blood* **111,** 1173–1181 (2008).
42. Raaijmakers, M. H. *et al.* Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature* **464,** 852–857 (2010).
This article demonstrates that perturbing specific mesenchymal populations in the bone marrow can result in pathological haematopoietic outcomes, including neoplasia.
43. Hosokawa, K. *et al.* Knockdown of N-cadherin suppresses the long-term engraftment of hematopoietic stem cells. *Blood* **116,** 554–563 (2010).
44. Hosokawa, K. *et al.* Cadherin-based adhesion is a potential target for niche manipulation to protect hematopoietic stem cells in adult bone marrow. *Cell Stem Cell* **6,** 194–198 (2010).
45. Wilson, A. *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev.* **18,** 2747–2763 (2004).
46. Li, P. & Zon, L. I. Resolving the controversy about N-cadherin and hematopoietic stem cells. *Cell Stem Cell* **6,** 199–202 (2010).
47. Ivanova, N. B. *et al.* A stem cell molecular signature. *Science* **298,** 601–604 (2002).
48. Seita, J. *et al.* Gene expression commons: an open platform for absolute gene expression profiling. *PLoS ONE* **7,** e40321 (2012).
49. Greenbaum, A. M., Revollo, L. D., Woloszynek, J. R., Civitelli, R. & Link, D. C. N-cadherin in osteolineage cells is not required for maintenance of hematopoietic stem cells. *Blood* **120,** 295–302 (2012).
50. Bromberg, O. *et al.* Osteoblastic N-cadherin is not required for microenvironmental support and regulation of hematopoietic stem and progenitor cells. *Blood* **120,** 303–313 (2012).
51. Guezguez, B. *et al.* Regional localization within the bone marrow influences the functional capacity of human HSCs. *Cell Stem Cell* **13,** 175–189 (2013).
52. Zhou, X. *et al.* Multiple functions of Osterix are required for bone growth and homeostasis in postnatal mice. *Proc. Natl Acad. Sci. USA* **107,** 12919–12924 (2010).
53. Chan, C. K. *et al.* Endochondral ossification is required for haematopoietic stem-cell niche formation. *Nature* **457,** 490–494 (2009).
54. Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131,** 324–336 (2007).
This article provides evidence for a mesenchymal skeletal stem cell that is capable of generating bone, secreting HSC niche factors and giving rise to bone marrow compartments that include HSC niches.
55. Ellis, S. L. *et al.* The relationship between bone, hemopoietic stem cells, and vasculature. *Blood* **118,** 1516–1524 (2011).
56. Adams, G. B. *et al.* Stem cell engraftment at the endosteal niche is specified by the calcium-sensing receptor. *Nature* **439,** 599–603 (2006).
57. Dai, J. C., He, P., Chen, X. & Greenfield, E. M. TNFα and PTH utilize distinct mechanisms to induce IL-6 and RANKL expression with markedly different kinetics. *Bone* **38,** 509–520 (2006).
58. Fleming, H. E. *et al.* Wnt signaling in the niche enforces hematopoietic stem cell quiescence and is necessary to preserve self-renewal *in vivo*. *Cell Stem Cell* **2,** 274–283 (2008).
59. Schaniel, C. *et al.* Wnt-inhibitory factor 1 dysregulation of the bone marrow niche exhausts hematopoietic stem cells. *Blood* **118,** 2420–2429 (2011).
60. Ferraro, F. *et al.* Diabetes impairs hematopoietic stem cell mobilization by altering niche function. *Sci. Transl. Med.* **3,** 104ra101 (2011).
61. Asada, N. *et al.* Matrix-embedded osteocytes regulate mobilization of hematopoietic stem/progenitor cells. *Cell Stem Cell* **12,** 737–747 (2013).
62. Greenbaum, A. *et al.* CXCL12 in early mesenchymal progenitors is required for haematopoietic stem-cell maintenance. *Nature* **495,** 227–230 (2013).
This paper is a systematic analysis of CXCL12-expressing cells in the bone marrow, demonstrating the specific role of primitive mesenchymal cells and endothelial cells in regulating HSC maintenance.
63. Méndez-Ferrer, S. *et al.* Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466,** 829–834 (2010).
This article reports that primitive mesenchymal cells residing perivascularly regulate HSC maintenance.
64. Ding, L., Saunders, T. L., Enikolopov, G. & Morrison, S. J. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* **481,** 457–462 (2012).
This paper is a systematic analysis of KitL expression in the bone marrow, demonstrating the requirement for endothelial and leptin-receptor-expressing perivascular cells in regulating HSC maintenance.
65. Omatsu, Y. *et al.* The essential functions of adipo-osteogenic progenitors as the

hematopoietic stem and progenitor cell niche. *Immunity* **33,** 387–399 (2010).
66. Morikawa, S. *et al*. Prospective identification, isolation, and systemic transplantation of multipotent mesenchymal stem cells in murine bone marrow. *J. Exp. Med*. **206,** 2483–2496 (2009).
67. Pinho, S. *et al*. PDGFRα and CD51 mark human Nestin$^+$ sphere-forming mesenchymal stem cells capable of hematopoietic progenitor cell expansion. *J. Exp. Med.* **210,** 1351–1367 (2013).
68. Tran, E. *et al.* Immune targeting of fibroblast activation protein triggers recognition of multipotent bone marrow stromal cells and cachexia. *J. Exp. Med.* **210,** 1125–1135 (2013).
69. Roberts, E. W. *et al*. Depletion of stromal cells expressing fibroblast activation protein-α from skeletal muscle and bone marrow results in cachexia and anemia. *J. Exp. Med*. **210,** 1137–1151 (2013).
70. Yao, L., Yokota, T., Xia, L., Kincade, P. W. & McEver, R. P. Bone marrow dysfunction in mice lacking the cytokine receptor gp130 in endothelial cells. *Blood* **106,** 4093–4101 (2005).
71. Li, W., Johnson, S. A., Shelley, W. C. & Yoder, M. C. Hematopoietic stem cell repopulating ability can be maintained *in vitro* by some primary endothelial cells. *Exp. Hematol*. **32,** 1226–1237 (2004).
72. Butler, J. M. *et al*. Endothelial cells are essential for the self-renewal and repopulation of Notch-dependent hematopoietic stem cells. *Cell Stem Cell* **6,** 251–264 (2010).
73. Kobayashi, H. *et al*. Angiocrine factors from Akt-activated endothelial cells balance self-renewal and differentiation of haematopoietic stem cells. *Nature Cell Biol*. **12,** 1046–1056 (2010).
74. Winkler, I. G. *et al*. Vascular niche E-selectin regulates hematopoietic stem cell dormancy, self renewal and chemoresistance. *Nature Med*. **18,** 1651–1657 (2012).
75. Broudy, V. C. Stem cell factor and hematopoiesis. *Blood* **90,** 1345–1364 (1997).
76. Czechowicz, A., Kraft, D., Weissman, I. L. & Bhattacharya, D. Efficient transplantation via antibody-based clearance of hematopoietic stem cell niches. *Science* **318,** 1296–1299 (2007).
77. Ogawa, M. *et al*. Expression and function of c-kit in hemopoietic progenitor cells. *J. Exp. Med.* **174,** 63–71 (1991).
78. Russell, E. S. Hereditary anemias of the mouse: a review for geneticists. *Adv. Genet.* **20,** 357–459 (1979).
79. Heissig, B. *et al*. Recruitment of stem and progenitor cells from the bone marrow niche requires MMP-9 mediated release of Kit-ligand. *Cell* **109,** 625–637 (2002).
80. Barker, J. E. Sl/Sld hematopoietic progenitors are deficient *in situ*. *Exp. Hematol*. **22,** 174–177 (1994).
81. Barker, J. E. Early transplantation to a normal microenvironment prevents the development of Steel hematopoietic stem cell defects. *Exp. Hematol*. **25,** 542–547 (1997).
82. Wolf, N. S. Dissecting the hematopoietic microenvironment. III. Evidence for a positive short range stimulus for cellular proliferation. *Cell Tissue Kinet.* **11,** 335–345 (1978).
83. Tzeng, Y. S. *et al*. Loss of Cxcl12/Sdf-1 in adult mice decreases the quiescent state of hematopoietic stem/progenitor cells and alters the pattern of hematopoietic regeneration after myelosuppression. *Blood* **117,** 429–439 (2011).
84. Nagasawa, T. *et al*. Defects of B-cell lymphopoiesis and bone-marrow myelopoiesis in mice lacking the CXC chemokine PBSF/SDF-1. *Nature* **382,** 635–638 (1996).
85. Petit, I. *et al*. G-CSF induces stem cell mobilization by decreasing bone marrow SDF-1 and up-regulating CXCR4. *Nature Immunol*. **3,** 687–694 (2002).
86. Ara, T. *et al*. Long-term hematopoietic stem cells require stromal cell-derived factor-1 for colonizing bone marrow during ontogeny. *Immunity* **19,** 257–267 (2003).
87. Zou, Y. R., Kottmann, A. H., Kuroda, M., Taniuchi, I. & Littman, D. R. Function of the chemokine receptor CXCR4 in haematopoiesis and in cerebellar development. *Nature* **393,** 595–599 (1998).
88. Ponomaryov, T. *et al*. Induction of the chemokine stromal-derived factor-1 following DNA damage improves human stem cell function. *J. Clin. Invest*. **106,** 1331–1339 (2000).
89. Dar, A. *et al*. Chemokine receptor CXCR4-dependent internalization and resecretion of functional chemokine SDF-1 by bone marrow endothelial and stromal cells. *Nature Immunol*. **6,** 1038–1046 (2005).
90. Hanoun, M. & Frenette, P. S. This niche is a maze; an amazing niche. *Cell Stem Cell* **12,** 391–392 (2013).
91. Katayama, Y. *et al*. Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow. *Cell* **124,** 407–421 (2006).
This paper reports evidence for nervous system involvement in regulating the bone marrow HSC niche.
92. Méndez-Ferrer, S., Lucas, D., Battista, M. & Frenette, P. S. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* **452,** 442–447 (2008).
This article demonstrates that neural circadian rhythms modulate HSC function.
93. Casanova-Acebes, M. *et al*. Rhythmic modulation of the hematopoietic niche through neutrophil clearance. *Cell* **153,** 1025–1035 (2013).
94. Chow, A. *et al*. Bone marrow CD169$^+$ macrophages promote the retention of hematopoietic stem and progenitor cells in the mesenchymal stem cell niche. *J. Exp. Med.* **208,** 261–271 (2011).
95. Winkler, I. G. *et al*. Bone marrow macrophages maintain hematopoietic stem cell (HSC) niches and their depletion mobilizes HSCs. *Blood* **116,** 4815–4828 (2010).
96. Yamazaki, S. *et al*. Nonmyelinating Schwann cells maintain hematopoietic stem cell hibernation in the bone marrow niche. *Cell* **147,** 1146–1158 (2011).
97. Kollet, O. *et al*. Osteoclasts degrade endosteal components and promote mobilization of hematopoietic progenitor cells. *Nature Med*. **12,** 657–664 (2006).
98. Mansour, A. *et al*. Osteoclasts promote the formation of hematopoietic stem cell niches in the bone marrow. *J. Exp. Med.* **209,** 537–549 (2012).
99. Nakada, D., Levi, B. P. & Morrison, S. J. Integrating physiological regulation with stem cell and tissue homeostasis. *Neuron* **70,** 703–718 (2011).
100. Qian, H. *et al*. Critical role of thrombopoietin in maintaining adult quiescent hematopoietic stem cells. *Cell Stem Cell* **1,** 671–684 (2007).
101. Yoshihara, H. *et al*. Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche. *Cell Stem Cell* **1,** 685–697 (2007).
102. Kimura, S., Roberts, A. W., Metcalf, D. & Alexander, W. S. Hematopoietic stem cell deficiencies in mice lacking c-Mpl, the receptor for thrombopoietin. *Proc. Natl Acad. Sci. USA* **95,** 1195–1200 (1998).
103. Kaushansky, K. Thrombopoietin and the hematopoietic stem cell. *Blood* **92,** 1–3 (1998).
104. Guerriero, A. *et al*. Thrombopoietin is synthesized by bone marrow stromal cells. *Blood* **90,** 3444–3455 (1997).
105. Sungaran, R., Markovic, B. & Chong, B. H. Localization and regulation of thrombopoietin mRNA expression in human kidney, liver, bone marrow, and spleen using *in situ* hybridization. *Blood* **89,** 101–107 (1997).
106. Kunisaki, Y. *et al*. Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* **502,** 637–643 (2013).
Histological characterization of subtypes of vascular structures and evidence that peri-arteriolar mesenchymal cells maintain HSC quiescence.
107. Parmar, K., Mauch, P., Vergilio, J. A., Sackstein, R. & Down, J. D. Distribution of hematopoietic stem cells in the bone marrow according to regional hypoxia. *Proc. Natl Acad. Sci. USA* **104,** 5431–5436 (2007).
108. Takubo, K. *et al*. Regulation of the HIF-1α level is essential for hematopoietic stem cells. *Cell Stem Cell* **7,** 391–402 (2010).
109. Lin, C. Direct measurement of local oxygen concentration in the bone marrow of live animals. *Nature* (in the press).
110. Visnjic, D. *et al*. Conditional ablation of the osteoblast lineage in Col2.3Δtk transgenic mice. *J. Bone Miner. Res*. **16,** 2222–2231 (2001).
111. Wu, J. Y. *et al*. Osteoblastic regulation of B lymphopoiesis is mediated by Gsα-dependent signaling pathways. *Proc. Natl Acad. Sci. USA* **105,** 16976–16981 (2008).
112. Nagasawa, T., Kikutani, H. & Kishimoto, T. Molecular cloning and structure of a pre-B-cell growth-stimulating factor. *Proc. Natl Acad. Sci. USA* **91,** 2305–2309 (1994).
113. Nie, Y., Han, Y. C. & Zou, Y. R. CXCR4 is required for the quiescence of primitive hematopoietic cells. *J. Exp. Med*. **205,** 777–783 (2008).
114. Chow, A. *et al*. CD169$^+$ macrophages provide a niche promoting erythropoiesis under homeostasis and stress. *Nature Med.* **19,** 429–436 (2013).
115. Himburg, H. A. *et al*. Pleiotrophin regulates the expansion and regeneration of hematopoietic stem cells. *Nature Med*. **16,** 475–482 (2010).
116. Himburg, H. A. *et al*. Pleiotrophin regulates the retention and self-renewal of hematopoietic stem cells in the bone marrow vascular niche. *Cell Rep*. **2,** 964–975 (2012).
117. Smith-Berdan, S. *et al*. Robo4 cooperates with CXCR4 to specify hematopoietic stem cell localization to bone marrow niches. *Cell Stem Cell* **8,** 72–83 (2011).
118. Smith-Berdan, S., Schepers, K., Ly, A., Passegue, E. & Forsberg, E. C. Dynamic expression of the Robo ligand Slit2 in bone marrow cell populations. *Cell Cycle* **11,** 675–682 (2012).
119. Nakamura-Ishizu, A. *et al*. Extracellular matrix protein tenascin-C is required in the bone marrow microenvironment primed for hematopoietic regeneration. *Blood* **119,** 5429–5437 (2012).
120. Stier, S. *et al*. Osteopontin is a hematopoietic stem cell niche component that negatively regulates stem cell pool size. *J. Exp. Med*. **201,** 1781–1791 (2005).
121. Nilsson, S. K. *et al*. Osteopontin, a key component of the hematopoietic stem cell niche and regulator of primitive hematopoietic progenitor cells. *Blood* **106,** 1232–1239 (2005).
122. Schepers, K. *et al*. Myeloproliferative neoplasia remodels the endosteal bone marrow niche into a self-reinforcing leukemic niche. *Cell Stem Cell* **13,** 285–299 (2013).
123. Walkley, C. R. *et al*. A microenvironment-induced myeloproliferative syndrome caused by retinoic acid receptor γ deficiency. *Cell* **129,** 1097–1110 (2007).
124. Walkley, C. R., Shea, J. M., Sims, N. A., Purton, L. E. & Orkin, S. H. Rb regulates interactions between hematopoietic stem cells and their bone marrow microenvironment. *Cell* **129,** 1081–1095 (2007).
125. Kiel, M. J. & Morrison, S. J. Uncertainty in the niches that maintain haematopoietic stem cells. *Nature Rev. Immunol*. **8,** 290–301 (2008).

Acknowledgements S.J.M. was supported by the National Heart, Lung and Blood Institute (HL097760), the Howard Hughes Medical Institute, and the Mary McDermott Cook Chair in Pediatric Genetics. D.T.S. was supported by the National Institutes of Health (HL044851, HL096372, EB014703) and the Gerald and Darlene Jordan Chair in Medicine. We apologize to authors whose work could not be cited because of space limitations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper at go.nature.com/pwfaf6. Correspondence should be addressed to S.J.M. (Sean.Morrison@UTSouthwestern.edu) and D.T.S. (David_Scadden@Harvard.edu).

REVIEW

doi:10.1038/nature12985

Mitochondrial form and function

Jonathan R. Friedman[1] & Jodi Nunnari[1]

Mitochondria are one of the major ancient endomembrane systems in eukaryotic cells. Owing to their ability to produce ATP through respiration, they became a driving force in evolution. As an essential step in the process of eukaryotic evolution, the size of the mitochondrial chromosome was drastically reduced, and the behaviour of mitochondria within eukaryotic cells radically changed. Recent advances have revealed how the organelle's behaviour has evolved to allow the accurate transmission of its genome and to become responsive to the needs of the cell and its own dysfunction.

Mitochondria arose around two billion years ago from the engulfment of an α-proteobacterium by a precursor of the modern eukaryotic cell[1]. Although mitochondria have maintained the double membrane character of their ancestors and the core of ATP production, their overall form and composition have been drastically altered, and they have acquired myriad additional functions within the cell. As part of the process of acquiring new functions during evolution, most of the genomic material of the α-proteobacterium progenitor was rapidly lost or transferred to the nuclear genome[2]. What remains in human cells is a small, approximately 16 kilobase, circular genome, which is present in cells in a vast excess of copies relative to nuclear chromosomes.

The human mitochondrial genome contains genetic coding information for 13 proteins, which are core constituents of the mitochondrial respiratory complexes I–IV that are embedded in the inner membrane. Functioning together with the Krebs' cycle in the matrix, the respiratory chain creates an electrochemical gradient through the coupled transfer of electrons to oxygen and the transport of protons from the matrix across the inner membrane into the intermembrane space. The electrochemical gradient powers the terminal complex V of the chain, ATP synthase, which is an ancient rotary turbine machine that catalyses the synthesis of most cellular ATP. The electrochemical potential is harnessed for additional crucial mitochondrial functions, such as buffering the signalling ion Ca^{2+} through uptake by a uniporter in the inner membrane[3,4]. A reduction in the electrochemical potential of mitochondria in cells has evolved as a read-out for mitochondrial functional status, which, as discussed later, creates signals to activate pathways that repair and/or eliminate defective mitochondria.

We know from a combination of proteomics, genomics and bioinformatics that modern day mitochondria are comprised of well over 1,000 proteins; the composition is plastic in nature, varying with and between species in response to cellular and tissue-specific organismal needs[5–7]. The origin of the mitochondrial proteome is a mixture of 'old' bacterial and 'new' eukaryotic-derived proteins[2]. For example, the mitochondrial DNA (mtDNA) replication and transcription machines have distinct evolutionary origins in bacteriophage[8–10], whereas the mitochondrial translational machinery has a clear evolutionary relationship to bacteria[11]. In addition to protein components, the mitochondrial genome encodes 22 transfer RNAs and 2 mitochondrial ribosome-coding RNAs, which are essential components of its own translational apparatus. Mitochondrial ribosome assembly in the mitochondrial matrix is a relatively complex and highly regulated process, which involves mitochondrial ribosome-coding RNA processing and maturation and the assembly of mitochondrial ribosomal proteins into small and large subunits[12]. However, only a fraction of mitochondrial ribosome proteins have identifiable homologues in bacteria[13]. The roles of mitochondrial specific ribosomal proteins are not understood, but these proteins are thought to have evolved to regulate the coordination of mitochondrial translation with extra-mitochondrial pathways in eukaryotic cells. Thus, like many mitochondrial machines, the ribosome is a mix of old and new innovations.

The nucleus-encoded proteins that make up most of the mitochondrial proteome are translated on cytosolic ribosomes and actively imported and sorted into mitochondrial sub-compartments by outer and inner membrane translocase machines in a manner that is dependent on the electrochemical potential[14,15]. Transcriptional, post-transcriptional and post-translational modes of regulation exist for nucleus-encoded mitochondrial proteins. In humans, transcriptional regulation of mitochondrial biogenesis occurs through the action of the PGC-1 family of co-activators, which respond to changes in nutrient status, such as NAD^+/NADH and AMP/ATP ratios (sensed through SIRT1 and AMPK, respectively), as well as environmental signals[16,17]. Combinatorial interactions between PGC-1 co-activators and specific transcription factors (NRF1, NRF2 and ERR) balance and specify the major functional pathways within mitochondria. Through the induction of nuclear genes that directly impinge on the maintenance of mtDNA, these interactions coordinate the nuclear and mitochondrial genomes[18]. Evidence in yeast suggests that nuclear-transcribed messenger RNAs encoding mitochondrial proteins are post-transcriptionally localized to the mitochondrial outer membrane in a highly regulated spatial and temporal manner, and coordinately translated[19,20]. Although the underlying molecular mechanisms of mRNA targeting to mitochondria are poorly understood, such pathways will probably be important in polarized cells such as neurons. Post-translational modifications, such as phosphorylation of mitochondrial import machinery components by cytosolic kinases, ultimately fine-tunes the proteome in response to metabolic cues[21].

Mutations in either mtDNA genes or nuclear genes that encode the mitochondrial proteins required for aerobic ATP production cause a diverse and often devastating array of human mitochondrial diseases, which can affect any organ in the body at any point during a person's life[22]. In addition, there is a high degree of clinical heterogeneity in mitochondrial diseases. Some of this heterogeneity can be explained by the fact that human cells can contain a variable ratio of mutated and wild-type mtDNA, a state called heteroplasmy. This seems to be the

[1]Department of Molecular and Cellular Biology, College of Biological Sciences, University of California, Davis, California 95616, USA.

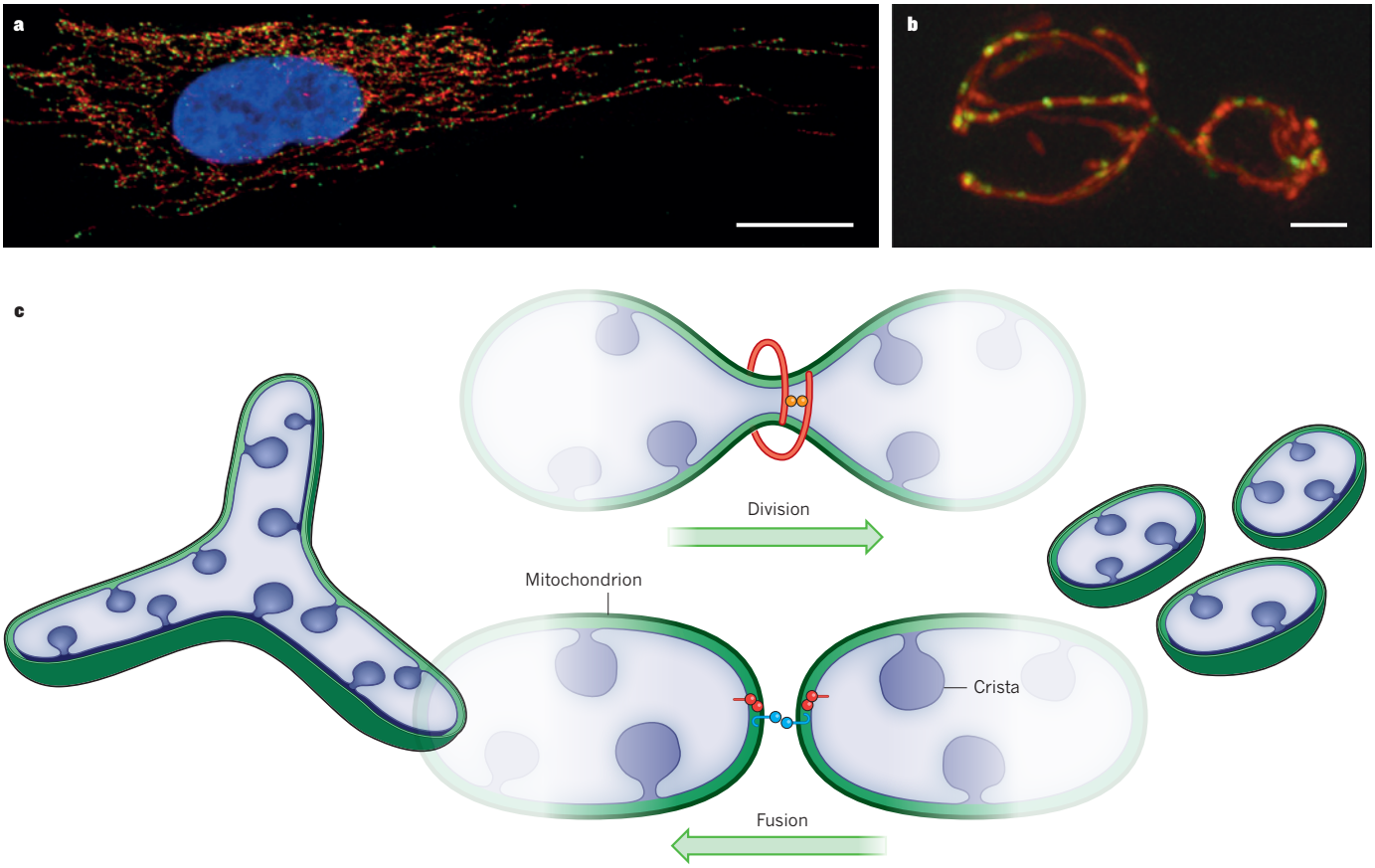


Figure 1 | The organization and distribution of mitochondria and mtDNA in higher eukaryotes. Mitochondrial organization is a conserved feature. **a**, mtDNA in a human fibroblast is packaged within nucleoids (green) distributed within tubular mitochondria (red) around the nucleus (blue). Scale bar, 20 microns. Adapted with permission from ref. 40. **b**, A similar distribution is seen in a yeast cell with nucleoids (green) within mitochondria (red). Scale bar, 2 microns. **c**, Mitochondrial copy number is controlled by the combined actions of mitochondrial division and fusion. Mitochondrial division is controlled by the assembly of a dynamin-related protein (DRP) on the outside of the organelle into a helical structure, which mediates scission through interactions across helical rungs (marked by orange circles). Mitochondrial fusion is controlled by interactions of outer and inner membrane fusion DRPs (blue and red circles, respectively) across mitochondrial membranes.

case for mtDNA mutations in protein-coding regions of the mitochondrial genome, in which an increase in mutant load gives rise to more severe disease phenotypes. However, disease heterogeneity that cannot be explained by heteroplasmy also exists when mutations are present in non-coding mitochondrial tRNAs. In addition, mutations in genes with a shared function, such as genes encoding subunits of complex I of the respiratory chain, lead to disease manifestations that are vastly different, such as optic nerve atrophy in adults or encephalopathy in infants[23,24]. Recent studies also point to a causal link between mutations in mtDNA and ageing, probably resulting from mtDNA-linked defects in somatic stem cells[25,26].

The role of mitochondria in disease has been expanded beyond the respiratory chain, as defects in additional mitochondrial functions and behaviours have been linked to cancer, metabolic disorders and neurodegenerative diseases, such as Alzheimer's, Parkinson's and Huntington's disease[22]. In general, however, our current understanding of the underlying relationship between mitochondrial phenotype and disease is poor and requires a better understanding of mitochondrial organization, as well as the connections mitochondria have with the nuclear genome and extra-mitochondrial pathways in different cell types and at the organismal level. To address this deficit, a renaissance in mitochondrial research has emerged, hastened by recent advances in genetics, systems-based approaches and our ability to visualize mitochondria at high temporal and spatial resolution. In this Review, we discuss how this renaissance is both redefining and extending our knowledge about mitochondrial behaviour and communication.

The mitochondrial chromosome

Given the importance of mtDNA-encoded genes for mitochondrial function, it is not surprising that there are dedicated mechanisms that actively control the structure and distribution of mitochondria and mtDNA, but in higher eukaryotes, these mechanisms are divergent from those of their ancestors[22]. Unlike bacteria, in most cell types, individual mitochondria do not exist; instead, they comprise a connected network containing multiple copies of the mitochondrial chromosome, forming a 'syncytium' (Fig. 1a, b). Like bacterial and nuclear chromosomes, mtDNA is highly compacted within the mitochondrial matrix, and consequently mtDNA–protein complexes can be visualized as punctate structures, termed nucleoids, within networks. The mechanism underlying mtDNA condensation was illuminated by researchers who solved the crystal structure of the most abundant mtDNA-associated protein in mammalian cells, mitochondrial transcription factor A (TFAM). The structure indicates that TFAM both binds and bends short stretches of mtDNA, forming loops that allow mtDNA packaging[27,28]. TFAM also plays a crucial part in mtDNA transcription, and its expression controls mtDNA copy number in cells, making it a central player in mtDNA maintenance and transmission[29,30].

Additional proteins that are crucial for mtDNA maintenance are localized at nucleoids. These proteins include replication and repair machinery, which, in humans, includes DNA polymerase γ and its accessory proteins, such as the replication helicase twinkle[31]. Mutations in genes encoding these and additional factors required for mtDNA maintenance cause a spectrum of human mitochondrial diseases, but at the molecular level, it is still not understood how these and the many

other identified mtDNA-associated proteins are assembled together and organized to build nucleoids[32]. This is despite the fact that proteomic inventories of mtDNA-associated proteins have been determined for many species[33–35]. In addition, both higher order nucleoid organization and the mode of mtDNA replication are variable between biological kingdoms, resulting in further complexity. These differences are a consequence of the composition of the genome and the nucleoid. For example, yeast possesses an active recombination machine similar to Rad52 recombination proteins, and replication is probably primarily recombination driven[36]. As a result, relative to the human mitochondrial genome, the larger 80 kilobase yeast genome is packaged in multiple copies in the nucleoid. By contrast, the mechanism of human mtDNA replication is recombination independent in most tissues and occurs through strand displacement[37,38]. Super-resolution imaging consistently indicates that human nucleoids contain a relatively small number of mtDNA molecules and thus are more solitary in nature[39,40]. These differences in organization and transmission modes of the mitochondrial chromosome greatly impact the segregation behaviour of mtDNA.

Modes of mtDNA segregation

The multi-copy nature of mtDNA means the mode of mtDNA transmission is viewed as stochastic or 'relaxed' in most cell types and thus is radically different from that of nuclear genes[41]. On an organismal level, in humans for example, mtDNA is inherited in a uniparental maternal manner, and paternal mtDNA is actively destroyed after fertilization[42,43]. Mitochondria in egg and sperm also have different functional states, shapes and cellular distributions, and these differences are probably important to confer fitness. In addition, from generation to generation mtDNA genotypes rapidly segregate, indicating that a 'bottleneck' exists to potentially eliminate severely defective mitochondria and/or mtDNA. In oocytes, the bottleneck is partly due to the manner in which mtDNA is replicated, as well as a consequence of mitochondrial organization. Oocyte mitochondria are organized into a transient structure called a Balbiani body, comprised of other organelles, such as the endoplasmic reticulum (ER) and Golgi, but the biogenesis of this structure is poorly understood[44]. During the reprogramming of fibroblasts into induced pluripotent stem (iPS) cells, heteroplasmic mtDNA genotypes also segregate through a bottleneck and mitochondria are organized into a Balbiani-like structure[45], suggesting that iPS cells could be a useful tool with which to study the cellular and molecular mechanisms of mtDNA genotype segregation during development.

Nuclear genes are replicated during a finite phase of the cell cycle and segregated by the concerted action of a microtubule-based spindle apparatus and an actin-based cytokinesis machine that work together to physically partition chromosomes into daughter cells. Bacterial cells also possess cell-cycle mechanisms to coordinate cell division with chromosome segregation through the placement of a tubulin-like FtsZ cell division machine. By contrast, the replication and segregation of mitochondrial chromosomes within most eukaryotes is not stringently coupled to the cell cycle, and at any given time, the replication of mtDNA occurs for only a subset of nucleoids in a given cell[46]. Bacterial cytoskeletal machinery has been lost during mitochondrial evolution, raising the question: what mechanisms are used to place division sites and segregate mtDNA? Such mechanisms will probably be important for understanding the cell- and tissue-specific mechanisms that underlie diseases linked to mtDNA mutations.

Dynamin-mediated mitochondrial dynamics

In higher eukaryotes, the segregation of mtDNA at the cellular level partly depends on continuous division and fusion events (Fig. 1c), whose rates are responsive to the needs of a particular cell type[47]. One fundamental role of mitochondrial fusion is to allow communication between organelles, perhaps to facilitate access to products of mtDNA expression[48–50]. Mitochondrial fusion also serves as a mechanism to buffer transient defects that arise in mitochondria[51]. Mitochondrial division antagonizes fusion-driven network assembly to facilitate mitochondrial distribution and transport by motor proteins on cytoskeletal networks to and from distal locations of demand[52]. A balance between division and fusion is important for the distribution and maintenance of mtDNA. Loss of mitochondrial fusion causes a normally connected network to fragment into multiple small mitochondria owing to unopposed division, and mtDNA is either completely or partially lost from cells with the associated severe defects in oxidative phosphorylation[50,53]. Attenuation of mitochondrial division causes mitochondria to elongate and form highly interconnected net-like structures, as well as causing defects in oxidative phosphorylation and mtDNA loss during cell division[54–57]. The link between mitochondrial dynamics and mtDNA transmission is consistent with the primary role of dynamics in the control of mitochondrial copy number. The more distributive nature of mitochondrial division coupled with opposing fusion has thus evolved to replace ancestral bacterial cytoskeletal machines.

Mitochondrial division and fusion events are mediated by the action of highly conserved dynamin-related proteins (DRPs) that, through their ability to self-assemble and hydrolyse GTP, facilitate membrane remodelling of diverse intracellular membranes[58]. Mitochondrial division is catalysed by a single DRP, DRP1 in mammals (Dnm1 in yeast). DRP1 and Dnm1 assemble through stalk domains into helical structures that wrap around the outer surface of mitochondria at constriction sites, whose diameters match the diameters of the division helix[59,60]. Interactions between the GTPase domains of division DRPs across the helical rungs catalyse the hydrolysis of GTP — an event thought to produce conformational changes that are transduced through the DRP helix to allow the coordinate scission of the outer and inner membranes[61–63]. Fusion of the mitochondrial outer and inner membranes requires the action of two evolutionarily distinct integral membrane DRPs, MFN1/MFN2 in mammals (Fzo1 in yeast) and OPA1 in mammals (Mgm1 in yeast), respectively[64]. Much less is known about how fusion DRPs work at a mechanistic level, although it is likely that interactions between the GTPase domains on opposing membranes are harnessed for membrane tethering and that self-assembly through the proposed stalk-like regions within a membrane is used for fusion.

The DRP family originates in bacteria, for which evidence suggests that members also function in membrane-linked processes[65]. However, DRPs' acquired roles as the primary machines that control mitochondrial copy number are a radical divergence from the bacterial FtsZ-dependent division machine, which works from the cytosolic face of the plasma membrane to mediate constriction and fission (Fig. 2). Insight into this transition to DRP-driven division comes from the division machines of primitive eukaryotic organelles from organisms such as the red algae *Cyanidioschyzon merolae*, and from endosymbiotic plastids and chloroplasts of most photosynthetic eukaryotes. These endosymbiotic organelles possess hybrid division machines, containing both internal FtsZ and external DRP assemblies[66,67]. The FtsZ machine functions in constriction and scission of the inner and outer membranes and in positioning the division site, whereas the DRP machine is recruited to the outer surface of the organelle — at the constriction site — and functions relatively late in the process to complete the scission of the outer membrane. As in the case for bacteria, FtsZ-dependent division site placement in these mitochondria and plastids is crucial for the transmission of organellar genomes.

ER-associated mitochondrial division

The loss of the FtsZ-like machine in higher eukaryotes raises questions of how and where division sites are placed in mitochondria and whether division site placement is important for mtDNA transmission. The answers have been partially addressed by the discovery that mitochondrial division site placement is dependent on a key inter-organellar interaction with the ER[68]. Before DRP1 recruitment to the mitochondrial outer membrane, ER tubules wrap around mitochondria and mark sites of mitochondrial division — a phenomenon, termed ER-associated mitochondrial division (ERMD), that has been conserved from yeast to humans. At these sites, mitochondria are constricted, and thus

geometric hotspots for the assembly of the division DRP helix are also created. In addition, the integral outer membrane DRP1 receptor and effector MFF[69] is recruited to sites of contact, which provides a spatial mark to link DRP1 recruitment to its activation and assembly into a division machine. Neither the mechanism underlying the generation of such an ER–mitochondrial microdomain nor of ER–associated mitochondrial constriction is understood. It is possible that the ER is able to directly alter mitochondrial membrane composition and/or morphology to facilitate the recruitment of factors localized on the outside and/or inside of mitochondria that promote mitochondrial constriction and division. In mammalian cells, the actin cytoskeleton has been implicated in ERMD, potentially through the ER-localized isoform of the formin INF2, raising the possibility that mitochondrial constriction during division is actin mediated[70].

ERMD must also involve a link or tether between the two organelles. The molecular basis of this link has recently been elucidated in yeast, and is mediated by a multiprotein complex termed the ER–mitochondrial encounter structure (ERMES). This structure forms a discrete and finite number of interfaces between the ER and mitochondria in cells[71,72]. In addition to marking sites of division, ERMES structures are tightly linked to a subset of nucleoids engaged in replicating mtDNA[46,73], potentially as components of a larger structure that spans multiple membranes. At sites of an ERMES complex, nucleoids segregate by an unknown mechanism and, in most cases, are distributed into both tips of divided mitochondria[72]. In this context, the ERMES complex has also been implicated as a bridge between mitochondria and the actin network, suggesting that it may link and coordinately drive nucleoid segregation, mitochondrial constriction during division, and mitochondrial distribution after division[74]. Thus, the process of ERMD and nucleoid segregation in yeast may fundamentally be related to the role of actin in ERMD in mammalian cells.

The distribution of daughter mitochondria following ERMD in yeast requires the highly conserved Miro GTPase Gem1 (ref. 72). Gem1 may function with ERMES to promote the resolution of daughter mitochondria by recruiting motility factors to mitochondrial tips after division. The metazoan Gem1 orthologues, MIRO1 and MIRO2 (hereafter referred to as Miro), also function in mitochondrial distribution. In this case, Miro proteins have been proposed to connect mitochondria to a member of the Milton/TRAK protein family of kinesin-1 adaptors to allow the microtubule-based transport of mitochondria[75,76]. However, although the Miro GTPase family is remarkably conserved, the mechanisms of mitochondrial transport are divergent in eukaryotes and, in yeast, mitochondrial motility is actin dependent. Thus, it is possible that the role of Gem1 and Miro GTPases in mitochondrial distribution may instead be to allow motility by directly regulating molecular tethers to disengage mitochondria from the ER at sites of division. In any case, Gem1-dependent distribution of daughter mitochondria after mitochondrial division serves as a cellular mechanism to coordinately distribute mitochondria and mtDNA.

Internal determinants of mitochondrial division

The observations discussed paint a compelling picture in which ERMD positions the division plane adjacent to mitochondrial nucleoids to bias their distribution into newly generated daughter mitochondria (Fig. 2). In mammalian cells, nucleoids are similarly localized at mitochondrial division sites and mitochondrial tips[77,78], and in the absence of the division DRP DRP1, nucleoids aggregate in clusters within hyperfused mitochondria[79]. This suggests that ERMD's role in nucleoid distribution is conserved, although the molecular identity of the ER–mitochondrial tether or tethers at division sites in mammalian cells is unknown. More

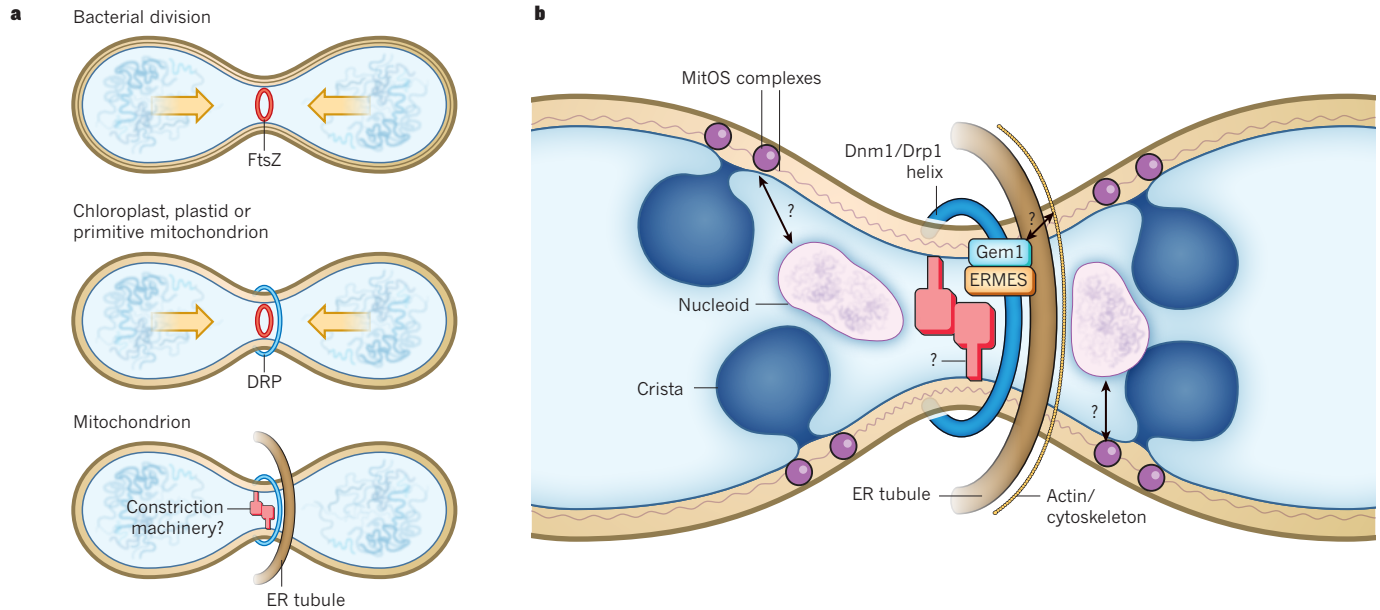


Figure 2 | Evolution of mitochondrial division site placement mechanisms. **a**, Roles of FtsZ and dynamin-related proteins (DRPs) in bacterial and endosymbiotic organelle division and division site placement. In the α-proteobacterial ancestor of mitochondria, an FtsZ ring is placed mid-cell by active mechanisms. The combined actions of the FtsZ-containing ring and cell-wall synthesis are essential for cell division. In chloroplasts and algal mitochondria, FtsZ-dependent placement (indicated by arrows) and division mechanisms on the inside of the organelle have been retained during evolution. However, in these organelles, DRPs also function on the cytosolic surface in organelle division, perhaps replacing the requirement for cell-wall synthesis in division. In yeast or animals, DRPs function on the cytosolic surface in organelle division. Before DRP recruitment, however, the endoplasmic reticulum (ER) is associated with division-site placement and constriction on the outside of the organelle, potentially replacing FtsZ-dependent placement and constriction. **b**, Molecular model for division site placement coupled to nucleoid segregation in yeast mitochondria. In yeast, the ER–mitochondria tethering complex, ER–mitochondrial encounter structure (ERMES), and the conserved Miro GTPase Gem1 are spatially and functionally linked to ER-associated mitochondrial division (ERMD). ERMD sites marked by these components are also spatially localized to a subset of nucleoids that are actively replicating, and these segregate before ERMD. Gem1 acts as a negative regulator of ERMES-dependent ER–mitochondria contacts and is required for the spatial resolution of newly generated mitochondria by ERMD, possibly through the localized recruitment of the actin cytoskeleton. Cytoskeletal components may also participate in ER-associated mitochondrial constriction before DRP recruitment. Nucleoid placement at sites of ERMD could be mediated by a mark inside the organelle formed by the scaffold MitOS complex. Mdm33 is a possible candidate for the internal membrane scission machine in yeast.

importantly, the question of what determines the placement of division sites still needs to be answered. Specifically, it is not known what determinants are required for the finite number and spatial localization of ER–mitochondria contact sites linked to nucleoid segregation. These questions are related to whether, in a manner analogous to bacterial FtsZ, there is a machine inside mitochondria that facilitates mitochondrial division. In yeast, an excellent candidate for the internal membrane scission machine is the inner membrane protein Mdm33, which possesses matrix-localized coiled-coil regions that could act in *trans* across inner membranes to mediate constriction[80].

Given the endosymbiotic origin of mitochondria, it is tempting to speculate that a spatial mark inside the organelles is used as a division-plane placement determinant. Evidence exists for an autonomous structure in mitochondria similar to the DNA-replicating replisome, which may function as such a mark. Specifically, nucleoid proteins required for mtDNA maintenance remain localized to discrete punctate structures within mitochondrial tubules in the absence of mtDNA in yeast and mammalian cells, suggesting that they have an intrinsic ability to organize into a structure in a mtDNA-independent manner[46,81]. In yeast, replisomes segregate within mitochondria and maintain their association with ERMES-marked ER–mitochondria contact sites, even in the absence of mitochondrial genomes[46].

Mitochondrial skeletal structures may also serve as internal spatial marks. Although bacterial-like cytoskeletal elements have apparently been lost, many scaffold-like structures exist within mitochondria, facilitating the formation of their complex external and internal structure. Mitochondrial scaffolds work together to create a higher-level organelle organization, which could encode spatial marks for nucleoid and/or division site placement. These scaffolds include the conserved prohibitin complex, which forms ring-like structures in the inner membrane that function together with mitochondrial lipids, such as cardiolipin and phosphatidylethanolamine, to organize inner membrane domains[82]. Another primary 'skeletal' element in mitochondria is the conserved multisubunit inner-membrane-associated complex MitOS (also called MICOS and MINOS)[83–85]. Evidence indicates that MitOS forms an extended heteromorphic structure that organizes and potentially shapes the mitochondrial inner membrane, which is differentiated into regions that are structurally, compositionally and functionally distinct. The region in close apposition to the outer membrane, termed the boundary region, facilitates lipid trafficking, mitochondrial protein import and respiratory complex assembly. Inner membrane cisternal invaginations called cristae house assembled respiratory complexes and have highly curved edges that are stabilized by the dimerization or multimerization of ATP synthase complexes[86]. Relatively narrow tubules, termed crista junctions, connect cristae to the boundary membrane and segregate soluble intermembrane-space components from the boundary regions. These junctions are restructured in apoptosis to promote the release of intermembrane-space-localized cell-death mediators into the cytosol during mitochondrial outer-membrane permeabilization (MOMP)[87]. Super-resolution imaging has revealed that mammalian nucleoids are tightly associated with inner membrane cristae[39]. Thus, the MitOS complex may also have a direct role in nucleoid positioning and/or may be part of a spatial mark that links the inside of mitochondria to the outside. Consistent with this possibility, in yeast, elements of the MitOS complex seem adjacent to nucleoids, and loss of an intact MitOS complex leads to nucleoid aggregation[88]. This complex also facilitates mitochondrial biogenesis by interacting with components of the import and sorting machineries in the outer mitochondrial membrane[85]. Thus, MitOS may have a more global function as a 'blueprint' of mitochondrial organization and in this capacity could act as a form-to-function integrator.

The ERMD microdomain as a systems integrator

Although the fundamental role of ER-associated division may be to control the distribution of mtDNA, evidence suggests that ERMD domains are harnessed for additional and diverse functions in cells and thus may also function as integrators. In addition to ERMD, the ERMES complex has been functionally linked to the biogenesis of outer membrane proteins and to lipid transport between the ER and mitochondria, which is crucial for the synthesis of essential mitochondrial lipids, such as phosphatidylethanolamine and cardiolipin[71,89]. ERMD domains thus might also monitor cellular status by facilitating communication between mitochondrial behaviours and cellular signalling pathways, for example, between mitochondrial division, fusion and cell death. Consistent with this possibility is the fact that DRP1 facilitates the recruitment and activation of the pro-apoptotic Bcl-2 protein BAX to the mitochondrial outer membrane to mediate MOMP[90]. Furthermore, ER-synthesized sphingolipids promote mitochondrial assembly of BAX and MOMP activation *in vitro*[91]. Conversely, mitochondrial fusion has a negative regulatory role in MOMP because cytosolic BAX promotes mitochondrial fusion *in vitro* through the DRP MFN2 (ref. 92), raising the possibility that fusion DRPs also regulate apoptosis through ERMD domains. It is possible that ERMD domains extend both into inner mitochondrial compartments and the ER lumen to integrate the functional status of both organelles as suggested by the regulation of MOMP by ER stress. Consistent with this, when ER stress occurs, the apoptotic regulator CDIP1 interacts with the ER protein BAP31, which subsequently leads to BAX assembly on mitochondria[93]. Thus, it will be interesting to determine whether sites of BAP31 and CDIP1 interaction coincide with sites of ERMD. The connection between the ER and mitochondria is substantiated by their roles in a shared set of diseases associated with altered mitochondrial dynamics. For example, proteins localized to ER–mitochondrial contacts have been implicated in Huntington's disease, optic atrophy and spinocerebellar ataxias[94]; and alteration of ER–mitochondrial contact has been described in Alzheimer's disease[95,96]. Thus, ERMD domain dysfunction may be a contributory factor in many diseases.

ERMD domains represent only one type of ER–mitochondrial contact. In yeast, for example, the ER is a component of two distinct Num1 and Mmr1 tethers that selectively position mitochondria at the cortex of mother and daughter cells, respectively, in a manner that is independent of ERMES and ERMD[97,98]. In addition, the fusion DRP MFN2, which is not essential to ERMD contact formation, has been proposed to act in mammalian cells as an ER–mitochondrial tether[68,99]. More work is needed, especially in mammalian cells, to understand the molecular basis of ER–mitochondria contacts and whether specialized contacts and tethers exist for different shared ER–mitochondrial functions, such as Ca^{2+} homeostasis, lipid biosynthesis, ERMD and for the ER–mitochondrial contacts implicated in autophagy[100,101]. This is an exciting area of mitochondrial biology that promises to be highly relevant to our understanding of the aetiology underlying diseases linked to mitochondrial dysfunction.

Coordination of diverse mitochondrial behaviours

Although mitochondrial division and fusion are major determinants of mitochondrial distribution, the behaviour of the mitochondrial network in cells is controlled by additional activities, such as tethering and motility. Neuronal cells are a prominent example of how these behavioural networks must work together responsively to maintain cellular function. Neurons are long, excitable cells that are highly compartmentalized and, for proper function, mitochondria must be appropriately distributed to serve the cells' different spatial and temporal demands. The demand for mitochondrial ATP production and Ca^{2+} buffering is especially high at axon terminals, which are dynamic structures that require the localized presence of mitochondria for synaptic transmission. Given that most mitochondrial biogenesis occurs in the soma of the neuron, active mechanisms are required to both transport and immobilize mitochondria at the distal synaptic terminals.

Insight into how these two processes are coordinated and function together to selectively target mitochondria to active synaptic terminals has come from recent studies on the neuron-specific protein syntaphilin, which binds specifically to the mitochondrial outer membrane and accumulates on immobilized axonal mitochondria localized to active terminals[102]. Mitochondria destined for axons are generated by mitochondrial

division in the soma and transported to the synapse along microtubules. The spatial link between division and nucleoids, and nucleoids and cytoskeletal elements, might, in this cell type, ensure that mitochondria destined for transport contain mtDNA. Syntaphilin functions as a brake, using at least two separate mechanisms. It binds directly to the microtubule-based kinesin motor KIF5 *in vitro* and inhibits its motor activity, suggesting that it converts KIF5 into a component of a static microtubule-dependent mitochondrial tether. Whether the ER plays a part in the biogenesis of the syntaphilin–KIF5 tether is an outstanding question. Syntaphilin also competes for KIF5 binding with the adaptor Milton/TRAK to facilitate tethering. Thus, there is extensive interplay between the motility and tethering machines to control mitochondrial distribution in an activity dependent and spatially specific manner.

Mitochondrial dynamics are probably also coordinated on a molecular level with the transport and tethering machineries in different cell types. In mammalian cells, mitochondria that lack the fusion DRP MFN2 show severe motility defects[103], similar to the motility defects observed in cells that lack Miro, and MFN2 has been reported to physically interact with Miro and Milton/TRAK[104]. This link between mitochondrial fusion and motility may be relevant for understanding why in humans, mutations in MFN2 and OPA1 cause the tissue-specific neurodegenerative diseases, Charcot Marie Tooth Type 2A (CMT2A) and dominant optic atrophy (DOA), respectively[22]. Indeed, a high frequency of mutations in mtDNA and nuclear genes that cause mitochondrial dysfunction selectively affect neurons and cause a diverse set of neurodegenerative diseases[32].

Mitochondrial quality control pathways

As well as being integrated with each other, mitochondrial behaviours are integrated with a battery of stress or quality-control pathways in cells that sense and respond to mitochondrial and cellular dysfunction (Fig. 3). Inside mitochondria, molecular chaperones and quality control proteases act together to promote the assembly of protein complexes comprised of mtDNA- and nucleus-encoded proteins as well as to monitor and degrade unfolded proteins[105]. An imbalance between nuclear and mitochondrial proteomes and/or an accumulation of unfolded mitochondrial proteins triggers a transcriptional response program in metazoans, termed the mitochondrial unfolded protein stress response pathway (UPRmt)[106–108]. The response is initiated by signals produced at the mitochondrial level that activate the transcription of nucleus-encoded mitochondrial chaperone genes, as well as additional genes to restore organelle homeostasis. The pathway has been characterized at the molecular level in *Caenorhabditis elegans*, for which both the mitochondrial-inner-membrane peptide transporter HAF1 and the bZip transcription factor ATFS1 are required for UPRmt signalling[109]. Recent evidence indicates that in healthy cells, ATFS1 is actively imported into the mitochondrial matrix, where it is constitutively degraded in healthy mitochondria[110]. Under conditions where the electron transport chain is disrupted, membrane potential is attenuated, and consequently the import efficiency of ATFS1 is decreased in a manner that is somehow dependent on HAF1. Extra-mitochondrial ATFS1 is stabilized and targeted to the nucleus, where it initiates a transcriptional response that increases the expression of mitochondrial chaperones and import machinery and remodels metabolism to rely less on respiration. Activation of the UPRmt in *C. elegans* is associated with an increase in lifespan, and recent evidence suggests that in mammals activation of this pathway also contributes to longevity, further implicating mitochondria as a crucial factor in ageing[106]. It remains to be determined whether the molecular mechanisms underlying the UPRmt are conserved in mammalian systems.

Additional mitochondrial stress-induced pathways are triggered by perturbations in electron transport chain function and/or reduction of membrane potential. The mitochondrial inner membrane fusion DRP OPA1 acts as a toggle between two pathways. In healthy cells, OPA1 processing occurs constitutively by the i-AAA protease YME1L to generate long transmembrane anchored and short, soluble isoforms, which are required for membrane fusion[111]. Reduction of mitochondrial membrane potential causes the long OPA1 isoforms to be converted into short forms by the metalloprotease OMA1, resulting in the inhibition of mitochondrial fusion and subsequent mitochondrial fragmentation[112,113]. This alteration serves to potentiate mitophagy and/or cell death. Conversely, long OPA1 isoforms are required for a different stress-induced response, termed mitochondrial hyperfusion. Mitochondrial hyperfusion mediates

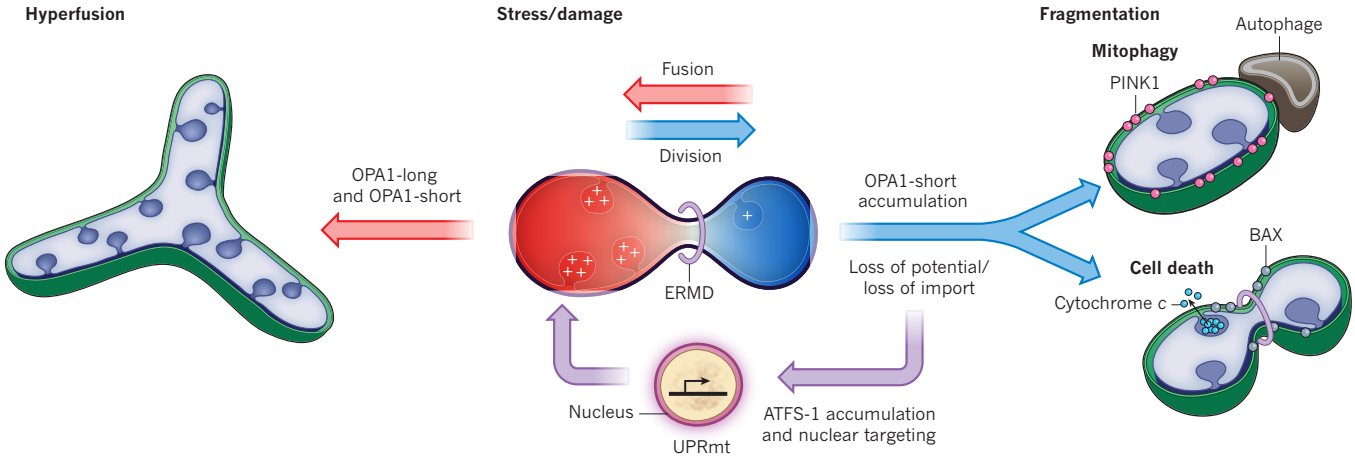


Figure 3 | Integration of mitochondrial stress response pathways and their coordination with mitochondrial shape. Several different mitochondrial pathways respond to stress or damage and are coordinated with mitochondrial dynamics. Mitochondria in healthy cells generate an electrochemical potential that serves in oxidative phosphorylation and drives the import of proteins into the organelle. Damage that leads to a loss (blue) of mitochondrial membrane potential can lead to a loss of protein import efficiency. In the unfolded protein stress response (UPRmt) pathway, loss of import leads to the accumulation of the transcription factor ATFS1 in the nucleus, activating a transcriptional mitochondrial repair and metabolic adaptation response. Loss of membrane potential also triggers the OMA1-dependent proteolysis of long isoforms of the inner membrane fusion DRP OPA1, which attenuates mitochondrial fusion and potentially increases ER-mediated mitochondrial division (ERMD), resulting in mitochondrial fragmentation. These fragmented mitochondria that have lost the ability to respire and import may also accumulate the PINK1 kinase, which triggers mitophagy. In addition, under these conditions, the ERMD domain may be altered to directly promote BAX oligomerization on the mitochondrial outer membrane, outer-membrane permeabilization and cytochrome *c* release, leading to cell death. Mitochondrial dysfunction and stresses such as starvation can trigger mitochondrial hyperfusion, which is dependent on maintenance of mitochondrial membrane potential (red) and the presence of both long and short OPA1 isoforms. Hyperfused mitochondria can transiently buffer the effects of respiratory chain dysfunction and do not enter the mitophagy pathway. The relationship between UPRmt and mitochondrial shape has not been explored.

the formation of a highly connected mitochondrial network and is thus thought to buffer the potentially deleterious effects of stresses, including those caused by ultraviolet irradiation and nutrient starvation. In the case of starvation, hyperfusion has been proposed to protect mitochondria from autophagic degradation or mitophagy through steric hinderance[114,115]. More recent work indicates that mitochondrial hyperfusion also serves as a homeostatic response to maintain ATP production in cases where complex IV of the electron transport chain is impaired[51]. The hyperfusion response, however, is transient and thus cannot buffer long-term defects in electron transport chain activity. A more terminal response to mitochondrial dysfunction is mitophagy, which is also triggered by a decrease in membrane-potential-driven protein import. In this pathway, the kinase PINK1 is imported into healthy mitochondria and constitutively degraded. A decrease in import triggered by mitochondrial dysfunction causes PINK1 to accumulate on the outer membrane, where it recruits the E3 ligase Parkin[116,117]. Among Parkin's targets for ubiquitination are the transport factor Miro and the mitochondrial fusion protein MFN2 (ref. 118–120). The proteasomal degradation of ubiquitinated mitochondrial outer membrane proteins is dependent on the AAA-ATPase p97 in a manner analogous to ER-associated degradation[120,121]. The Parkin-dependent degradation of factors involved in mitochondrial motility and fusion enhances the selectivity of removing defective mitochondria by autophagy. In addition to specifying defective mitochondria for degradation, an *in vivo* study in *Drosophila* suggests that the PINK1–Parkin pathway may also be capable of selectively targeting respiratory complexes for degradation[122]. In support of this idea, the selective targeting of complex I for degradation has also been described in cell culture models[45], but the mechanisms underlying this phenomenon are currently unknown.

Mitochondrial stress pathways probably have important roles in disease manifestations of mitochondrial dysfunction and, as such, could highlight promising therapeutic targets. In the case of the PINK1–Parkin mitophagy pathway, mutations in each gene are linked to genetically inherited Parkinson's disease[123,124], implying the pathway is relevant. However, most of the stress pathways described above, particularly those caused by a loss of membrane potential, have been characterized under conditions of acute extreme stress, raising the question of their physiological relevance[125]. In addition, despite these numerous stress pathways, mtDNA mutations are able to accumulate, particularly in differentiated post-mitotic cells, at the expense of functional wild-type mtDNA, leading to disease states. Thus, future work should focus on developing animal models that mimic diseases associated with mitochondrial dysfunction to assess the physiological contributions of these pathways. In addition, continued basic biological approaches are necessary to assess how the cell appropriately senses, and differentially activates and coordinates these pathways with each other and with other signalling pathways, such as those involved in cell death (Fig. 3). For example, it is unclear how cells appropriately integrate the UPRmt and mitophagy pathways, which are both regulated at the basic level of import. It is also unknown whether OPA1 can function as a molecular integrator of stresses or simply as a modulator of mitochondrial shape or how OPA1-dependent stress pathways are coordinated with UPRmt and mitophagy.

The emerging picture of mitochondria is that of 'super-organized' structural domains for building an organelle whose behaviour is wired to be responsive to cellular needs, as well as its own dysfunction. The combined use of system-based approaches, with super-resolution microscopy and new genetic tools will allow us to understand the molecular basis of mitochondrial structure in detail. Exactly how mitochondrial super-organization is constructed will address the fundamental question of whether primary determinants of organization originate from the inside of the organelle and are intimately tied to the most ancient feature — the genome. ■

Received 23 September; accepted 22 November 2013.

1. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–934 (2010).
2. Gabaldón, T. & Huynen, M. A. Shaping the mitochondrial proteome. *Biochim. Biophys. Acta* **1659,** 212–220 (2004).
3. Baughman, J. M. *et al*. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476,** 341–345 (2011).
4. De Stefani, D., Raffaello, A., Teardo, E., Szabo, I. & Rizzuto, R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476,** 336–340 (2011).
5. Pagliarini, D. J. *et al*. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–123 (2008).
6. Sickmann, A. *et al*. The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl Acad. Sci. USA* **100,** 13207–13212 (2003).
7. Forner, F., Foster, L. J., Campanaro, S., Valle, G. & Mann, M. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol. Cell. Proteomics* **5,** 608–619 (2006).
8. Lecrenier, N., Van Der Bruggen, P. & Foury, F. Mitochondrial DNA polymerases from yeast to man: a new family of polymerases. *Gene* **185,** 147–152 (1997).
9. Stumpf, J. D. & Copeland, W. C. Mitochondrial DNA replication and disease: insights from DNA polymerase γ mutations. Cellular and molecular life sciences. *Cell. Mol. Life Sci.* **68,** 219–233 (2011).
10. Tiranti, V. *et al*. Identification of the gene encoding the human mitochondrial RNA polymerase (h-mtRPOL) by cyberscreening of the expressed sequence tags database. *Hum. Mol. Genet.* **6,** 615–625 (1997).
11. Christian, B. E. & Spremulli, L. L. Mechanism of protein biosynthesis in mammalian mitochondria. *Biochim. Biophys. Acta* **1819,** 1035–1054 (2012).
12. Fung, S., Nishimura, T., Sasarman, F. & Shoubridge, E. A. The conserved interaction of C7orf30 with MRPL14 promotes biogenesis of the mitochondrial large ribosomal subunit and mitochondrial translation. *Mol. Biol. Cell* **24,** 184–193 (2013).
13. Sharma, M. R. *et al*. Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell* **115,** 97–108 (2003).
14. Neupert, W. & Herrmann, J. M. Translocation of proteins into mitochondria. *Annu. Rev. Biochem.* **76,** 723–749 (2007).
15. Schmidt, O., Pfanner, N. & Meisinger, C. Mitochondrial protein import: from proteomics to functional mechanisms. *Nature Rev. Mol. Cell Biol.* **11,** 655–667 (2010).
16. Jäger, S., Handschin, C., St-Pierre, J. & Spiegelman, B. M. AMP-activated protein kinase (AMPK) action in skeletal muscle via direct phosphorylation of PGC-1α. *Proc. Natl Acad. Sci. USA* **104,** 12017–12022 (2007).
17. Jeninga, E. H., Schoonjans, K. & Auwerx, J. Reversible acetylation of PGC-1: connecting energy sensors and effectors to guarantee metabolic flexibility. *Oncogene* **29,** 4617–4624 (2010).
18. Scarpulla, R. C., Vega, R. B. & Kelly, D. P. Transcriptional integration of mitochondrial biogenesis. *Trends Endocrinol. Metab.* **23,** 459–466 (2012).
19. Gadir, N., Haim-Vilmovsky, L., Kraut-Cohen, J. & Gerst, J. E. Localization of mRNAs coding for mitochondrial proteins in the yeast *Saccharomyces cerevisiae*. *RNA* **17,** 1551–1565 (2011).
20. Garcia, M. *et al*. Mitochondria-associated yeast mRNAs and the biogenesis of molecular complexes. *Mol. Biol. Cell* **18,** 362–368 (2007).
21. Schmidt, O. *et al*. Regulation of mitochondrial protein import by cytosolic kinases. *Cell* **144,** 227–239 (2011).
22. Nunnari, J. & Suomalainen, A. Mitochondria: in sickness and in health. *Cell* **148,** 1145–1159 (2012).
23. Wallace, D. C. *et al*. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242,** 1427–1430 (1988).
24. Morris, A. A. *et al*. Deficiency of respiratory chain complex I is a common cause of Leigh disease. *Ann. Neurol.* **40,** 25–30 (1996).
25. Ross, J. M. *et al*. Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. *Nature* **501,** 412–415 (2013).
26. Ahlqvist, K. J. *et al*. Somatic progenitor cell vulnerability to mitochondrial DNA mutagenesis underlies progeroid phenotypes in Polg mutator mice. *Cell Metab.* **15,** 100–109 (2012).
27. Ngo, H. B., Kaiser, J. T. & Chan, D. C. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nature Struct. Mol. Biol.* **18,** 1290–1296 (2011).
28. Rubio-Cosials, A. *et al*. Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. *Nature Struct. Mol. Biol.* **18,** 1281–1289 (2011).
29. Ekstrand, M. I. *et al*. Mitochondrial transcription factor A regulates mtDNA copy number in mammals. *Hum. Mol. Genet.* **13,** 935–944 (2004).
30. Shi, Y. *et al*. Mammalian transcription factor A is a core component of the mitochondrial transcription machinery. *Proc. Natl Acad. Sci. USA* **109,** 16510–16515 (2012).
31. Bogenhagen, D. F. Mitochondrial DNA nucleoid structure. *Biochim. Biophys. Acta* **1819,** 914–920 (2012).
32. Copeland, W. C. Defects in mitochondrial DNA replication and human disease. *Crit. Rev. Biochem. Mol. Biol.* **47,** 64–74 (2012).
33. Bogenhagen, D. F., Wang, Y., Shen, E. L. & Kobayashi, R. Protein components of mitochondrial DNA nucleoids in higher eukaryotes. *Mol. Cell. Proteomics* **2,** 1205–1216 (2003).
34. Kaufman, B. A. *et al. In organello* formaldehyde crosslinking of proteins to mtDNA: identification of bifunctional proteins. *Proc. Natl Acad. Sci. USA* **97,** 7772–7777 (2000).
35. He, J. *et al*. Mitochondrial nucleoid interacting proteins support mitochondrial protein synthesis. *Nucleic Acids Res.* **40,** 6109–6121 (2012).

36. Mbantenkhu, M. *et al*. Mgm101 is a Rad52-related protein required for mitochondrial DNA recombination. *J. Biol. Chem.* **286,** 42360–42370 (2011).
37. Holt, I. J. & Reyes, A. Human mitochondrial DNA replication. *Cold Spring Harb. Perspect. Biol.* **4,** a012971 (2012).
38. Brown, T. A., Tkachuk, A. N. & Clayton, D. A. Native R-loops persist throughout the mouse mitochondrial DNA genome. *J. Biol. Chem*. **283,** 36743–36751 (2008).
39. Brown, T. A. *et al*. Superresolution fluorescence imaging of mitochondrial nucleoids reveals their spatial range, limits, and membrane interaction. *Mol. Cell. Biol.* **31,** 4994–5010 (2011).
40. Kukat, C. *et al*. Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. *Proc. Natl Acad. Sci. USA* **108,** 13534–13539 (2011).
41. Birky, C. W., Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu. Rev. Genet.* **35,** 125–148 (2001).
42. Al Rawi, S. *et al*. Postfertilization autophagy of sperm organelles prevents paternal mitochondrial DNA transmission. *Science* **334,** 1144–1147 (2011).
43. Sato, M. & Sato, K. Degradation of paternal mitochondria by fertilization-triggered autophagy in *C. elegans* embryos. *Science* **334,** 1141–1144 (2011).
44. Pepling, M. E., Wilhelm, J. E., O'Hara, A. L., Gephardt, G. W. & Spradling, A. C. Mouse oocytes within germ cell cysts and primordial follicles contain a Balbiani body. *Proc. Natl Acad. Sci. USA* **104,** 187–192 (2007).
45. Hämäläinen, R. H. *et al*. Tissue- and cell-type-specific manifestations of heteroplasmic mtDNA 3243A>G mutation in human induced pluripotent stem cell-derived disease model. *Proc. Natl Acad. Sci. USA* **110,** E3622–E3630 (2013).
This paper showed that during reprogramming of heteroplasmic fibroblasts derived from mitochondrial disease patients, mutant and wild-type mitochondrial genomes segregate through a bottleneck towards a homoplasmic state.
46. Meeusen, S. & Nunnari, J. Evidence for a two membrane-spanning autonomous mitochondrial DNA replisome. *J. Cell Biol.* **163,** 503–510 (2003).
47. Hoppins, S., Lackner, L. & Nunnari, J. The machines that divide and fuse mitochondria. *Annu. Rev. Biochem.* **76,** 751–780 (2007).
48. Chen, H., McCaffery, J. M. & Chan, D. C. Mitochondrial fusion protects against neurodegeneration in the cerebellum. *Cell* **130,** 548–562 (2007).
49. Chen, H. *et al*. Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. *Cell* **141,** 280–289 (2010).
50. Hermann, G. J. *et al*. Mitochondrial fusion in yeast requires the transmembrane GTPase Fzo1p. *J. Cell Biol.* **143,** 359–373 (1998).
51. Rolland, S. G. *et al.* Impaired complex IV activity in response to loss of LRPPRC function can be compensated by mitochondrial hyperfusion. *Proc. Natl Acad. Sci. USA* **110,** E2967–E2976 (2013).
This paper demonstrates that mitochondria can undergo hyperfusion and temporarily maintain ATP production to compensate for a reduction of complex IV activity due to loss of the RNA-binding protein LRPPRC.
52. Verstreken, P. *et al*. Synaptic mitochondria are critical for mobilization of reserve pool vesicles at *Drosophila* neuromuscular junctions. *Neuron* **47,** 365–378 (2005).
53. Chen, H., Chomyn, A. & Chan, D. C. Disruption of fusion results in mitochondrial heterogeneity and dysfunction. *J. Biol. Chem*. **280,** 26185–26192 (2005).
54. Ishihara, N. *et al*. Mitochondrial fission factor Drp1 is essential for embryonic development and synapse formation in mice. *Nature Cell Biol*. **11,** 958–966 (2009).
55. Parone, P. A. *et al*. Preventing mitochondrial fission impairs mitochondrial function and leads to loss of mitochondrial DNA. *PLoS ONE* **3,** e3257 (2008).
56. Wakabayashi, J. *et al*. The dynamin-related GTPase Drp1 is required for embryonic and brain development in mice. *J. Cell Biol*. **186,** 805–816 (2009).
57. Hanekamp, T. *et al*. Maintenance of mitochondrial morphology is linked to maintenance of the mitochondrial genome in *Saccharomyces cerevisiae*. *Genetics* **162,** 1147–1156 (2002).
58. Faelber, K. *et al*. Oligomerization of dynamin superfamily proteins in health and disease. *Prog. Mol. Biol. Transl. Sci*. **117,** 411–443 (2013).
59. Ingerman, E. *et al*. Dnm1 forms spirals that are structurally tailored to fit mitochondria. *J. Cell Biol.* **170,** 1021–1027 (2005).
60. Labrousse, A. M., Zappaterra, M. D., Rube, D. A. & van der Bliek, A. M. *C. elegans* dynamin-related protein DRP-1 controls severing of the mitochondrial outer membrane. *Mol. Cell* **4,** 815–826 (1999).
61. Ford, M. G., Jenni, S. & Nunnari, J. The crystal structure of dynamin. *Nature* **477,** 561–566 (2011).
62. Fröhlich, C. *et al*. Structural insights into oligomerization and mitochondrial remodelling of dynamin 1-like protein. *EMBO J*. **32,** 1280–1292 (2013).
63. Faelber, K. *et al*. Crystal structure of nucleotide-free dynamin. *Nature* **477,** 556–560 (2011).
64. Meeusen, S., McCaffery, J. M. & Nunnari, J. Mitochondrial fusion intermediates revealed *in vitro*. *Science* **305,** 1747–1752 (2004).
65. Low, H. H., Sachse, C., Amos, L. A. & Lowe, J. Structure of a bacterial dynamin-like protein lipid tube provides a mechanism for assembly and membrane curving. *Cell* **139,** 1342–1352 (2009).
66. Osteryoung, K. W. & Nunnari, J. The division of endosymbiotic organelles. *Science* **302,** 1698–1704 (2003).
67. Nishida, K. *et al*. Dynamic recruitment of dynamin for final mitochondrial severance in a primitive red alga. *Proc. Natl Acad. Sci. USA* **100,** 2146–2151 (2003).
68. Friedman, J. R. *et al*. ER tubules mark sites of mitochondrial division. *Science* **334,** 358–362 (2011).
69. Otera, H. *et al*. Mff is an essential factor for mitochondrial recruitment of Drp1 during mitochondrial fission in mammalian cells. *J. Cell Biol*. **191,** 1141–1158 (2010).
70. Korobova, F., Ramabhadran, V. & Higgs, H. N. An actin-dependent step in mitochondrial fission mediated by the ER-associated formin INF2. *Science* **339,** 464–467 (2013).
71. Kornmann, B. *et al*. An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science* **325,** 477–481 (2009).
72. Murley, A. *et al*. ER-associated mitochondrial division links the distribution of mitochondria and mitochondrial DNA in yeast. *eLife* **2,** e00422 (2013).
In yeast, ERMD serves to segregate mitochondrial genomes into tips of newly divided mitochondria, and the conserved Miro GTPase Gem1 may spatially resolve ER–mitochondrial contacts post-division.
73. Hobbs, A. E., Srinivasan, M., McCaffery, J. M. & Jensen, R. E. Mmm1p, a mitochondrial outer membrane protein, is connected to mitochondrial DNA (mtDNA) nucleoids and required for mtDNA stability. *J. Cell Biol*. **152,** 401–410 (2001).
74. Boldogh, I. R. *et al*. A protein complex containing Mdm10p, Mdm12p, and Mmm1p links mitochondrial membranes and DNA to the cytoskeleton-based segregation machinery. *Mol. Biol. Cell* **14,** 4618–4627 (2003).
75. Glater, E. E., Megeath, L. J., Stowers, R. S. & Schwarz, T. L. Axonal transport of mitochondria requires milton to recruit kinesin heavy chain and is light chain independent. *J. Cell Biol.* **173,** 545–557 (2006).
76. Fransson, S., Ruusala, A. & Aspenstrom, P. The atypical Rho GTPases Miro-1 and Miro-2 have essential roles in mitochondrial trafficking. *Biochem. Biophys. Res. Commun*. **344,** 500–510 (2006).
77. Iborra, F. J., Kimura, H. & Cook, P. R. The functional organization of mitochondrial genomes in human cells. *BMC Biol*. **2,** 9 (2004).
78. Garrido, N. *et al*. Composition and dynamics of human mitochondrial nucleoids. *Mol. Biol. Cell* **14,** 1583–1596 (2003).
79. Ban-Ishihara, R., Ishihara, T., Sasaki, N., Mihara, K. & Ishihara, N. Dynamics of nucleoid structure regulated by mitochondrial fission contributes to cristae reformation and release of cytochrome c. *Proc. Natl Acad. Sci. USA* **110,** 11863–11868 (2013).
80. Messerschmitt, M. *et al*. The inner membrane protein Mdm33 controls mitochondrial morphology in yeast. *J. Cell Biol*. **160,** 553–564 (2003).
81. Spelbrink, J. N. *et al*. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nature Genet*. **28,** 223–231 (2001).
82. Osman, C., Voelker, D. R. & Langer, T. Making heads or tails of phospholipids in mitochondria. *J. Cell Biol*. **192,** 7–16 (2011).
83. Hoppins, S. *et al*. A mitochondrial-focused genetic interaction map reveals a scaffold-like complex required for inner membrane organization in mitochondria. *J. Cell Biol*. **195,** 323–340 (2011).
84. Harner, M. *et al*. The mitochondrial contact site complex, a determinant of mitochondrial architecture. *EMBO J*. **30,** 4356–4370 (2011).
85. von der Malsburg, K. *et al.* Dual role of mitofilin in mitochondrial membrane organization and protein biogenesis. *Dev. Cell* **21,** 694–707 (2011).
86. Davies, K. M., Anselmi, C., Wittig, I., Faraldo-Gomez, J. D. & Kuhlbrandt, W. Structure of the yeast F_1F_o-ATP synthase dimer and its role in shaping the mitochondrial cristae. *Proc. Natl Acad. Sci. USA* **109,** 13602–13607 (2012).
87. Frezza, C. *et al*. OPA1 controls apoptotic cristae remodeling independently from mitochondrial fusion. *Cell* **126,** 177–189 (2006).
88. Itoh, K., Tamura, Y., Iijima, M. & Sesaki, H. Effects of Fcj1-Mos1 and mitochondrial division on aggregation of mitochondrial DNA nucleoids and organelle morphology. *Mol. Biol. Cell* **24,** 1842–1851 (2013).
89. Voss, C., Lahiri, S., Young, B. P., Loewen, C. J. & Prinz, W. A. ER-shaping proteins facilitate lipid exchange between the ER and mitochondria in *S. cerevisiae*. *J. Cell Sci*. **125,** 4791–4799 (2012).
90. Montessuit, S. *et al*. Membrane remodeling induced by the dynamin-related protein Drp1 stimulates Bax oligomerization. *Cell* **142,** 889–901 (2010).
91. Chipuk, J. E. *et al*. Sphingolipid metabolism cooperates with BAK and BAX to promote the mitochondrial pathway of apoptosis. *Cell* **148,** 988–1000 (2012).
92. Hoppins, S. *et al.* The soluble form of Bax regulates mitochondrial fusion via MFN2 homotypic complexes. *Mol. Cell* **41,** 150–160 (2011).
93. Namba, T. *et al*. CDIP1–BAP31 complex transduces apoptotic signals from endoplasmic reticulum to mitochondria under endoplasmic reticulum stress. *Cell Rep.* **5,** 331–339 (2013).
94. Schon, E. A. & Przedborski, S. Mitochondria: the next (neurode) generation. *Neuron* **70,** 1033–1053 (2011).
95. Area-Gomez, E. *et al.* Upregulated function of mitochondria-associated ER membranes in Alzheimer disease. *EMBO J*. **31,** 4106–4123 (2012).
96. Hedskog, L. *et al*. Modulation of the endoplasmic reticulum-mitochondria interface in Alzheimer's disease and related models. *Proc. Natl Acad. Sci. USA* **110,** 7916–7921 (2013).
97. Lackner, L. L., Ping, H., Graef, M., Murley, A. & Nunnari, J. Endoplasmic reticulum-associated mitochondria-cortex tether functions in the distribution and inheritance of mitochondria. *Proc. Natl Acad. Sci. USA* **110,** E458–E467 (2013).
98. Swayne, T. C. *et al*. Role for cER and Mmr1p in anchorage of mitochondria at sites of polarized surface growth in budding yeast. *Curr. Biol.* **21,** 1994–1999 (2011).
99. de Brito, O. M. & Scorrano, L. Mitofusin 2 tethers endoplasmic reticulum to mitochondria. *Nature* **456,** 605–610 (2008).
100. Rowland, A. A. & Voeltz, G. K. Endoplasmic reticulum-mitochondria contacts: function of the junction. *Nature Rev. Mol. Cell Biol*. **13,** 607–625 (2012).

101. Hamasaki, M. *et al.* Autophagosomes form at ER-mitochondria contact sites. *Nature* **495,** 389–393 (2013).
102. Chen, Y. & Sheng, Z. H. Kinesin-1-syntaphilin coupling mediates activity-dependent regulation of axonal mitochondrial transport. *J. Cell Biol*. **202,** 351–364 (2013).
The authors show that the protein syntaphilin can regulate mitochondrial position in neurons by acting as a molecular brake through its binding to the microtubule motor Kif5.
103. Baloh, R. H., Schmidt, R. E., Pestronk, A. & Milbrandt, J. Altered axonal mitochondrial transport in the pathogenesis of Charcot-Marie-Tooth disease from mitofusin 2 mutations. *J. Neurosci*. **27,** 422–430 (2007).
104. Misko, A., Jiang, S., Wegorzewska, I., Milbrandt, J. & Baloh, R. H. Mitofusin 2 is necessary for transport of axonal mitochondria and interacts with the Miro/Milton complex. *J. Neurosci*. **30,** 4232–4240 (2010).
105. Baker, M. J., Tatsuta, T. & Langer, T. Quality control of mitochondrial proteostasis. *Cold Spring Harb. Perspect. Biol.* **3,** a007559 (2011).
106. Houtkooper, R. H. *et al*. Mitonuclear protein imbalance as a conserved longevity mechanism. *Nature* **497,** 451–457 (2013).
107. Zhao, Q. *et al*. A mitochondrial specific stress response in mammalian cells. *EMBO J*. **21,** 4411–4419 (2002).
108. Martinus, R. D. *et al*. Selective induction of mitochondrial chaperones in response to loss of the mitochondrial genome. *Eur. J. Biochem*. **240,** 98–103 (1996).
109. Haynes, C. M., Yang, Y., Blais, S. P., Neubert, T. A. & Ron, D. The matrix peptide exporter HAF-1 signals a mitochondrial UPR by activating the transcription factor ZC376.7 in *C. elegans*. *Mol. Cell* **37,** 529–540 (2010).
110. Nargund, A. M., Pellegrino, M. W., Fiorese, C. J., Baker, B. M. & Haynes, C. M. Mitochondrial import efficiency of ATFS-1 regulates mitochondrial UPR activation. *Science* **337,** 587–590 (2012).
This work demonstrated that in *C. elegans*, the transcription factor ATFS-1 senses and modulates a response to mitochondrial stress through its targeting to either the mitochondrial matrix or the nucleus.
111. Griparic, L., Kanazawa, T. & van der Bliek, A. M. Regulation of the mitochondrial dynamin-like protein Opa1 by proteolytic cleavage. *J. Cell Biol.* **178,** 757–764 (2007).
112. Head, B., Griparic, L., Amiri, M., Gandre-Babbe, S. & van der Bliek, A. M. Inducible proteolytic inactivation of OPA1 mediated by the OMA1 protease in mammalian cells. *J. Cell Biol*. **187,** 959–966 (2009).
113. Ehses, S. *et al*. Regulation of OPA1 processing and mitochondrial fusion by m-AAA protease isoenzymes and OMA1. *J. Cell Biol*. **187,** 1023–1036 (2009).
114. Rambold, A. S., Kostelecky, B., Elia, N. & Lippincott-Schwartz, J. Tubular network formation protects mitochondria from autophagosomal degradation during nutrient starvation. *Proc. Natl Acad. Sci. USA* **108,** 10190–10195 (2011).
115. Gomes, L. C., Di Benedetto, G. & Scorrano, L. During autophagy mitochondria elongate, are spared from degradation and sustain cell viability. *Nature Cell Biol.* **13,** 589–598 (2011).
116. Matsuda, N. *et al*. PINK1 stabilized by mitochondrial depolarization recruits Parkin to damaged mitochondria and activates latent Parkin for mitophagy. *J. Cell Biol*. **189,** 211–221 (2010).
117. Narendra, D. P. *et al*. PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. *PLoS Biol.* **8,** e1000298 (2010).
118. Wang, X. *et al*. PINK1 and Parkin target Miro for phosphorylation and degradation to arrest mitochondrial motility. *Cell* **147,** 893–906 (2011).
119. Chan, N. C. *et al*. Broad activation of the ubiquitin-proteasome system by Parkin is critical for mitophagy. *Hum. Mol. Genet.* **20,** 1726–1737 (2011).
120. Tanaka, A. *et al*. Proteasome and p97 mediate mitophagy and degradation of mitofusins induced by Parkin. *J. Cell Biol.* **191,** 1367–1380 (2010).
121. Xu, S., Peng, G., Wang, Y., Fang, S. & Karbowski, M. The AAA-ATPase p97 is essential for outer mitochondrial membrane protein turnover. *Mol. Biol. Cell* **22,** 291–300 (2011).
122. Vincow, E. S. *et al*. The PINK1-Parkin pathway promotes both mitophagy and selective respiratory chain turnover *in vivo*. *Proc. Natl Acad. Sci. USA* **110,** 6400–6405 (2013).
The authors utilized proteomics of *Drosophila* Pink1 and Parkin mutants to show that respiratory complex components are selectively turned over compared with other mitochondrial proteins during mitophagy.
123. Kitada, T. *et al*. Mutations in the *parkin* gene cause autosomal recessive juvenile parkinsonism. *Nature* **392,** 605–608 (1998).
124. Valente, E. M. *et al*. Hereditary early-onset Parkinson's disease caused by mutations in *PINK1*. *Science* **304,** 1158–1160 (2004).
125. Sterky, F. H., Lee, S., Wibom, R., Olson, L. & Larsson, N. G. Impaired mitochondrial transport and Parkin-independent degeneration of respiratory chain-deficient dopamine neurons *in vivo*. *Proc. Natl Acad. Sci. USA* **108,** 12937–12942 (2011).

Acknowledgements We thank members of the Nunnari lab for helpful discussions and comments. We also thank K. Osteryoung and S. Lewis for helpful discussions. J.N. is supported by NIH grants R01GM062942, R01GM097432 and R01GM106019. J.F. is supported by a fellowship from the Jane Coffin Childs Memorial Fund for Medical Research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/v7zzkf. Correspondence should be addressed to J.N. (jmnunnari@ucdavis.edu).

REVIEW

doi:10.1038/nature12986

The multilayered complexity of ceRNA crosstalk and competition

Yvonne Tay[1], John Rinn[2,3,4] & Pier Paolo Pandolfi[1]

Recent reports have described an intricate interplay among diverse RNA species, including protein-coding messenger RNAs and non-coding RNAs such as long non-coding RNAs, pseudogenes and circular RNAs. These RNA transcripts act as competing endogenous RNAs (ceRNAs) or natural microRNA sponges — they communicate with and co-regulate each other by competing for binding to shared microRNAs, a family of small non-coding RNAs that are important post-transcriptional regulators of gene expression. Understanding this novel RNA crosstalk will lead to significant insight into gene regulatory networks and have implications in human development and disease.

In recent years, numerous studies have documented pervasive transcription across 70–90% of the human genome. This was particularly surprising because less than 2% of the total genome encodes protein-coding genes, suggesting that non-coding RNAs represent most of the human transcriptome. Recent reports indicate that aside from around 21,000 protein-coding genes, the human transcriptome includes about 9,000 small RNAs, about 10,000–32,000 long non-coding RNAs (lncRNAs) and around 11,000 pseudogenes[1,2]. Non-coding transcripts can generally be divided into two major classes on the basis of their size. Small non-coding RNAs have been relatively well characterized, and include transfer RNAs, which are involved in translation of messenger RNAs; microRNAs (miRNAs) and small-interfering RNAs, which are implicated in post-transcriptional RNA silencing; small nuclear RNAs, which are involved in splicing; small nucleolar RNAs, which are implicated in ribosomal RNA modification; PIWI-interacting RNAs, which are involved in transposon repression; and transcription initiation RNAs, promoter upstream transcripts and promoter-associated small RNAs, which may be involved in transcription regulation. lncRNAs can vary in length from 200 nucleotides to 100 kilobases, and have been implicated in a diverse range of biological processes from pluripotency to immune responses[3]. One of the best-studied and most dramatic examples is *XIST*, a single RNA gene that can recruit chromatin-modifying complexes to inactivate an entire chromosome during dosage compensation[4]. However, although thousands of lncRNAs have been identified in the past decade, only a small number have been functionally characterized.

The importance of the non-coding transcriptome has become increasingly clear in recent years — comparative genomic analysis has demonstrated a significant difference in genome utilization among species (for example, the protein-coding genome constitutes almost the entire genome of unicellular yeast, but only 2% of mammalian genomes)[5], and the non-coding transcriptome is often dysregulated in cancer[6]. These observations suggest that the non-coding transcriptome is of crucial importance in determining the greater complexity of higher eukaryotes and in disease pathogenesis[7,8]. Functionalizing the non-coding space will undoubtedly lead to important insight about basic physiology and disease progression.

Although many reviews have focused on the regulatory mechanisms and functions of lncRNAs[3,9], there are many additional implications for the pervasive transcription observed in mammalian genomes. In this Review, we focus on the emerging roles of RNA–RNA crosstalk, which include new layers of gene regulation that involve interactions between diverse RNA species. A classic example of RNA–RNA interactions involves the post-transcriptional regulation of RNA transcripts by miRNAs. As our knowledge of the transcriptome space has expanded, it has become increasingly clear that numerous miRNA-binding sites exist on a wide variety of RNA transcripts, leading to the hypothesis that all RNA transcripts that contain miRNA-binding sites can communicate with and regulate each other by competing specifically for shared miRNAs, thus acting as competing endogenous RNAs (ceRNAs)[10–12]. miRNA competition thus extends beyond the non-coding transcriptome and potentially confers an additional non-protein-coding function to protein-coding mRNAs. Although it has been proposed that ceRNA crosstalk may be limited to a small subset of transcripts, owing to factors such as miRNA abundance, ceRNA abundance and subcellular localization[10,13], the discovery of functional ceRNA regulation in diverse species — including viruses, plants, mice and humans — by multiple independent groups suggests that it may represent a widespread layer of gene regulation[14–18]. We discuss literature describing the effect of miRNA competition on the regulation of both non-coding and coding RNAs, additional factors that may affect ceRNA activity and potential directions for future studies, as well as the implications of miRNA competition for development and disease.

RNA crosstalk in competitive endogenous networks

Competition between various molecular species to bind to a specific molecular target has been described in multiple contexts, and includes DNA–protein, RNA–protein, RNA–RNA and protein–protein crosstalk. This Review focuses on crosstalk involving RNA–RNA interactions and an additional regulatory layer of RNA–protein interactions.

Protein–RNA competition

RNA-binding proteins (RBPs) are key regulators of multiple post-transcriptional events, including RNA splicing, stability, transport and translation[19]. Various RBPs may compete to bind to specific target transcripts: the RBP HuR, which generally stabilizes target transcripts; and AUF1, which generally leads to the rapid degradation of target transcripts, have been shown to compete for common target binding sites[20,21]; and competition between HuR and wild-type

[1]Cancer Research Institute, Beth Israel Deaconess Cancer Center, Department of Medicine and Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. [2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA. [3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [4]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142, USA.

TTP for binding to the *HuR* (also known as *Elavl1*) transcript has been implicated in HuR regulation and cytoplasmic localization[22]. Conversely, target transcripts may compete to bind to specific RBPs: competition between *GAP43* and *β-actin* mRNAs for binding to the K-homology (KH)-domain RBP ZBP1 has been shown to affect their exonal localization[23].

miRNA–RNA competition

In addition to the protein-coding dimension, molecular competition also extends to regulatory networks comprised exclusively of RNAs, suggesting that sequence competition represents a universal and prevalent form of gene regulation. Two intriguing new players, which add further complexity to RNA crosstalk, are miRNAs and lncRNAs. Experimental evidence has confirmed that competition for miRNAs, small non-coding regulators of gene expression, plays an integral part in the regulation of both lncRNAs[14–16,24,25] and mRNAs[17,18,26,27]. We summarize the background and experimental evidence that supports this competitive RNA crosstalk and discuss potential refinements and directions for future analyses.

Artificial miRNA sponges as miRNA competitors

Several years before the discovery of naturally occurring miRNA sponges, or ceRNAs, various groups described the use of artificial miRNA sponges as effective miRNA inhibitors[28–30]. These sponges are usually expressed from strong promoters, contain multiple binding sites for an miRNA of interest and have been shown to derepress miRNA targets at least as effectively as chemically modified antisense oligonucleotides[30]. The efficacy of artificial sponges has been demonstrated for multiple miRNAs both *in vitro* and *in vivo*[30–34]. Artificial miRNA sponges constructed with multiple binding sites for different miRNAs may also be used to study the effect of several miRNAs simultaneously[30,32].

Intriguingly, although sponges with perfectly complementary miRNA-binding sites have been shown to be effective[29,30,33], 'bulged sponges', which include a central bulge and hence bind miRNAs with imperfect complementarity, have been demonstrated to sequester miRNAs with greater efficacy[30–33,35,36]. This may be partly due to the fact that, unlike perfectly complementary targets, imperfect targets are not immediately degraded and are thus able to reduce miRNA bioavailability until the mRNA is destabilized by other factors[31,35]. These artificial miRNA sponges will not only be invaluable tools for miRNA loss-of-function studies *in vitro* and *in vivo*, but may also provide a new platform for RNA-based therapeutic applications.

Natural miRNA sponges as ceRNAs

It has been proposed that naturally occurring protein-coding and non-coding RNA transcripts can act as endogenous miRNA sponges, or ceRNAs[11,12,15]. ceRNAs communicate with, and co-regulate, each other by competing to bind to shared miRNAs, thereby titrating miRNA availability. Because ceRNA crosstalk can be deciphered bioinformatically and experimentally, the identification and analysis of ceRNA interactions may allow the systematic functionalization of the non-coding transcriptome, as well as a non-protein-coding function to be attributed to mRNA transcripts, which may be complementary or even distinct from their protein-coding function[11].

Non-coding RNAs as ceRNAs

Increasing experimental evidence supports the hypothesis that multiple non-coding RNA species, including small non-coding RNAs, pseudogenes, lncRNAs and circular RNAs (circRNAs) may possess ceRNA activity.

Plant and viral non-coding ceRNAs

One of the first examples of endogenous non-coding miRNA sponges was described in plants[15] (Table 1). The non-coding RNA *IPS1* from *Arabidopsis thaliana* has been reported to alter the stability of *PHO2* mRNA by sequestering the phosphate starvation-induced miRNA miR-399. In addition, although most miRNA targets in plants are cleaved owing to almost perfect miRNA complementarity, the *IPS1* motif contained a mismatched loop at the miRNA cleavage site that abolished transcript cleavage and resulted in effective miR-399 sequestration. Generation of a cleavable *IPS1* variant eliminated its inhibitory activity on miR-399. This was consistent with the observations from artificial sponge constructs suggesting that imperfectly complementary 'bulged sponges' sequester miRNAs more effectively than perfectly complementary miRNA sponges.

Intriguingly, the primate virus *Herpesvirus saimiri* has been shown to use a non-coding ceRNA to control host-cell gene expression[14]. *H. saimiri*-transformed T cells express several viral non-coding RNAs of unknown function called *H. saimiri* U-rich RNAs (HSURs). One of these non-coding RNAs, HSUR1, has been found to contain miR-27-binding sites and direct miR-27 degradation in a sequence-specific and binding-dependent manner. The expression of HSUR1 and HSUR2 was shown to correlate with an upregulation of FOXO1 levels, a validated miR-27 target, suggesting that perturbation of miRNA expression by HSURs was able to control host gene expression.

Table 1 | List of validated non-coding competing endogenous RNAs

Non-coding RNA species	Competing endogenous RNAs		Shared microRNAs	Organism	Reference
Small non-coding RNA	*HSUR1*	*FOXO1*	miR-27a	*Herpesvirus saimiri*	14
Long non-coding RNA	*IPS1*	*PHO2*	miR-399	*Arabidopsis thaliana*	15
	HULC	*PRKACB*	miR-372	*Homo sapiens*	25
	linc-MD1	*MAML1* *MEF2C*	miR-133 miR-135	*Mus musculus* and *Homo sapiens*	24
	linc-RoR	*NANOG* *OCT4* *SOX2*	miR-145	*Homo sapiens*	43
	PTCSC3		miR-574-5p	*Homo sapiens*	41
	H19		Let-7 family	*Mus musculus* and *Homo sapiens*	42
Pseudogene	*PTENP1*	*PTEN*	miR-17, miR-19, miR-21, miR-26 and miR-214 families	*Homo sapiens*	16
	KRAS1P	*KRAS*	Let-7 family		
	Pbcas4	*BCAS4*	miR-185	*Mus musculus* and *Homo sapiens*	39
Circular RNA	*CDR1as/ciRS-7*		miR-7	*Danio rerio*, *Mus musculus* and *Homo sapiens*	45, 46
	Sry		miR-138	*Mus musculus* and *Homo sapiens*	45

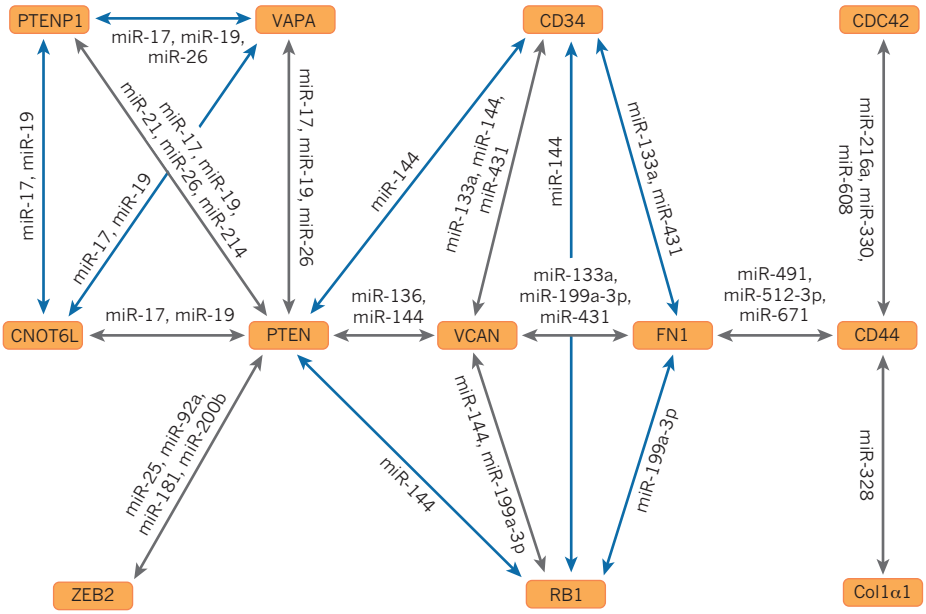


Figure 1 | *PTEN* competing endogenous RNA (ceRNA) network. Validated interactions (grey arrows) between ceRNAs have been reported, shown here with the respective validated microRNAs (miRNAs). Studies have also identified potential ceRNA interactions (blue arrows) based on the respective validated miRNAs. Validated *PTEN* ceRNAs described in ref. 26 are not included because the specific miRNAs were not identified.

Pseudogene ceRNAs

In humans, the non-coding *PTENP1* pseudogene has been reported to regulate levels of its cognate gene, the tumour suppressor *PTEN*, by competing for shared miRNAs[16] (Fig. 1). *PTENP1* has been shown to act as a decoy for *PTEN*-targeting miRNAs, to possess tumour suppressive activity and to be selectively lost in human cancers. An additional dimension to *PTENP1*-mediated ceRNA regulation was recently outlined[37], whereby two antisense RNAs encoded by the *PTENP1* locus α and β were characterized. The β isoform was shown to interact with the PTENP1 transcript through an RNA–RNA pairing interaction, affecting its stability and ceRNA activity[37].

In addition, overexpression of the *KRAS* pseudogene *KRAS1P* 3′ untranslated region (UTR) has been reported to increase KRAS transcript abundance and accelerate cell proliferation. The *KRAS1P* locus was amplified in various human cancers, and *KRAS1P* transcript levels positively correlated with *KRAS* transcript levels in prostate cancer. Taken together, these findings were particularly relevant because they attributed a new function to transcribed pseudogenes, lncRNAs, which have been considered to be biologically inconsequential owing to their inability to be translated into functional proteins. Although transcribed pseudogenes may be expressed at much lower levels than their cognate genes, this is counterbalanced by their high degree of shared sequence homology, which results in the conservation of multiple miRNA-binding sites and allows them to compete for the binding of many shared miRNAs simultaneously. Furthermore, it has been suggested that RNA transcripts that contain premature stop codons, such as pseudogenes, may be subjected to nonsense-mediated mRNA decay[38]. This rapid turnover may conceivably lead to their low abundance, as well as the increased degradation of bound miRNAs, enhancing their effectiveness as miRNA sponges.

Another example of a pseudogene ceRNA was described by Marques *et al.*, who focused their attention on 'unitary pseudogenes', which retain protein-coding capability in the human lineage but lose protein-coding function in the rodent lineage. This allowed the RNA-mediated function of RNAs to be disentangled from their ancestral protein-coding function. The mouse *Pbcas4* transcript was validated as a conserved ceRNA for human *BCAS4,* which downregulated expression levels of the protein-coding transcripts *Bcl2*, *Il17rd*, *Pnpla3*, *Shisa7* and *Tapbp* by acting as a miR-185 decoy[39]. The preservation of miRNA response elements (MREs) and thus miRNA sponge function of this unitary pseudogene after the loss of its protein-coding function lends support to the hypothesis that ceRNA interactions represent a conserved and biologically relevant mechanism of post-transcriptional gene regulation, which confers an additional non-protein-coding role to protein-coding transcripts.

lncRNA ceRNAs

The complexity and diversity of potential ceRNA interactions have scaled exponentially with the identification of more than 10,000 lncRNAs. Examples are already emerging of lncRNAs as competitive platforms for both miRNAs and mRNAs. HuR-mediated binding of the miRNA *let-7* to the long intergenic non-coding RNA *lincRNA-p21* was shown to lead to decreased *lincRNA-p21* transcript levels and a concomitant increase in the translation of the *lincRNA-p21*-associated mRNAs *JUNB* and *CTNNB1* (ref. 40).

Many lncRNAs are known to have low abundance and/or nuclear localization[9], which may adversely affect their effectiveness as ceRNAs in steady state conditions. However, it has been shown that thousands of lncRNAs possess cell type-, tissue type-, developmental stage- and disease-specific expression patterns and localization[9], suggesting that individual lncRNAs may be potent natural miRNA sponges in certain settings. Two such examples of lncRNA ceRNAs that have emerged, underscoring the importance of these miRNA–lncRNA competitive interactions, are *HULC* and *PTCSC3*.

The lncRNA *HULC* has been identified as one of the most significantly upregulated transcripts in hepatocellular carcinoma (HCC)[25]. The *HULC* transcript was found to contain miR-372-binding sites, and its overexpression could reduce miR-372 expression and activity in the liver cancer cell line Hep3B. *HULC*'s ceRNA activity forms part of an intricate autoregulatory loop: miR-372 inhibition by *HULC* reduces translational repression of its target transcript *PRKACB*, inducing cAMP response element binding protein (CREB) phosphorylation and enhancing CREB-dependent HULC upregulation in liver cancer. Significantly, this observed interaction extended beyond *in vitro* cell-line data to human clinical samples: miR-372 expression was downregulated in all HCC tissue samples compared with corresponding adjacent normal tissue, consistent with increased *HULC* expression.

In contrast to *HULC,* which is upregulated in HCC, *PTCSC3* was found to be dramatically downregulated in thyroid cancers. *PTCSC3* transfection has been shown to lead to a significant decrease in

expression levels of the oncogenic miR-574-5p, as well as growth inhibition, cell-cycle arrest and increased apoptosis in three thyroid cancer cell lines[41]. These two studies suggest that both the disease-specific upregulation and downregulation of individual lncRNAs, and the resultant changes in ceRNA-mediated interactions, may have profound effects in pathophysiological conditions.

In addition to their roles in various cancers, lncRNA ceRNAs have been implicated in human development. For example, a muscle-specific lncRNA, *lincMD1*, which is activated on myoblast differentiation and controls muscle differentiation in human and mouse myoblasts through its ceRNA activity, has been identified[24]. *lincMD1* sequestered miR-133 and miR-135, effectively regulating the expression of *MAML1* and *MEF2C* mRNAs, respectively. Importantly, when *lincMD1* levels were in excess, the repression of *MAML1* and *MEF2C* could be titrated by increasing expression levels of the respective miRNAs, confirming that the observed regulation was due to direct competition for those specific miRNAs. As *MAML1* and *MEF2C* encode transcription factors known to activate muscle-specific gene expression, these data suggest that ceRNA regulation is of crucial importance in myogenic differentiation. In another study, developmentally regulated lncRNA H19 was found to contain binding sites for the let-7 miRNA family, and thus acted as an effective ceRNA for the very abundant let-7 miRNA, hence modulating the expression of other let-7 target transcripts including Dicer and Hmga2 (ref. 42). The observation that let-7 overexpression could recapitulate the precocious muscle differentiation caused by *H19* knockdown provides further evidence of the physiological relevance of ceRNA interactions in this setting.

Sequestration of miRNAs has also been shown to be of functional importance in pluripotent embryonic stem (ES) cells. The lncRNA *linc-RoR*, which is abundantly expressed in human ES cells and downregulated during differentiation, has been shown to share regulatory miRNAs with OCT4, SOX2 and NANOG, which are core transcription factors that are essential for ES cell self-renewal[43]. *linc-RoR* effectively sequestered miR-145, protecting *OCT4*, *SOX2* and *NANOG* transcripts from miR-145-mediated suppression. Significantly, this regulation was abolished by the introduction of mutations in the two miR-145-binding sites in *linc-RoR*, providing further evidence that the observed effect was miR-145 dependent. As *linc-RoR* transcription is directly regulated by these core transcription factors, these results implicate it in a regulatory feedback loop in ES cells and suggest that its ceRNA function is essential for ES cell pluripotency and self-renewal.

circRNA ceRNAs

About 20 years ago, the predominant transcript of the testis-determining gene *Sry* in mouse testis was found to be circular[44], although the physiological relevance of these RNA circles remained elusive. The recent discovery of competitive RNA–RNA interactions coupled with the extensive complementarity of circRNAs to their linear mRNA counterparts has raised the possibility that these RNA circles may have an integral role in regulatory RNA networks. Recently, a circRNA called *CDR1as* (also known as *ciRS-7*)[45,46] was identified. This highly stable circRNA contains more than 60 conserved binding sites for miR-7 and hence acts as an effective miR-7 sponge that affects miR-7 target gene activity. In zebrafish, its expression impaired midbrain development in a manner analogous to *miR-7* knockdown[46]. However, this is not an isolated example of circRNAs with ceRNA activity, *Sry* has also been validated as a miR-138 sponge[45]. These studies represent the first functional analysis of circRNAs.

Recent bioinformatic and experimental analyses have identified thousands of circRNAs in the mammalian transcriptome, suggesting that circRNAs may in fact represent a new class of ceRNA regulators[45–47]. Importantly, owing to their high expression levels and increased stability, circRNAs with ceRNAs activity may be exceptionally effective modulators of the crosstalk between linear ceRNAs[48]. circRNAs such as *CDR1as* that contain multiple binding sites for the same miRNA may represent a mechanism for sequestering more abundant miRNAs, which would be significantly less susceptible to titration by transcripts that contain only one miRNA-binding site. In addition to their ceRNA function, it is possible that circRNAs may also bind and sequester RBPs, base pair with other RNAs or even produce proteins[49].

Taken together, these studies suggest that the analysis of ceRNA crosstalk may provide invaluable insight into the function of diverse species of non-coding RNAs, including lncRNAs, pseudogenes and circRNAs. Furthermore, because any RNA transcript that contains miRNA-binding sites can sequester miRNAs, ceRNA crosstalk also extends beyond the non-coding space to confer a non-protein-coding function to mRNAs.

mRNAs as ceRNAs

The gene with perhaps the most extensively characterized ceRNA network is the important tumour suppressor *PTEN*. Aside from the non-coding pseudogene *PTENP1* (discussed earlier), the *PTEN* ceRNA network includes multiple protein-coding transcripts (Fig. 1, Table 2).

A combined computational and experimental approach has been used to identify *CNOT6L* and *VAPA* as protein-coding transcripts that regulate *PTEN* transcript and protein expression in a Dicer-dependent manner, antagonize downstream PI(3)K signalling and possess growth- and tumour-suppressive properties[17]. These genes were also coexpressed with *PTEN* mRNA in several human cancers and displayed copy number loss in colon cancer. The study demonstrated that previously uncharacterized transcripts could be functionalized, partly through the identification of their ceRNA interactors, and presented a framework

Table 2 | List of validated protein-coding competing endogenous RNAs

Competing endogenous RNAs		Shared microRNAs	Organism	Reference
PTEN	*CNOT6L*	miR-17 and miR-19 families	*Homo sapiens*	17
	VAPA	miR-17, miR-19 and miR-26 families		
	ZEB2	miR-25, miR-92a, miR-181 and miR-200b	*Homo sapiens* and *Mus musculus*	18
	ABHD13, *CCDC6*, *CTBP2*, *DCLK1*, *DKK1*, *HIAT1*, *HIF1A*, *KLF6*, *LRCH1*, *NRAS*, *RB1*, *TAF5* and *TNKS2*		*Homo sapiens*	26
PTEN	*VCAN*	miR-136 and miR-144	*Homo sapiens* and *Mus musculus*	50
Rb1	*VCAN*	miR-144 and miR-199a-3p		
CD34		miR-133a, miR-144 and miR-431		51
FN1		miR-133a, miR-199a-3p and miR-431		27,51
FN1	*CD44*	miR-491, miR-512-3p and miR-671	*Homo sapiens*	53
Col1a1	*CD44*	miR-328		
CDC42		miR-216a, miR-330 and miR-608		52
HMGA2	*TGFBR3*	Let-7 family	*Homo sapiens* and *Mus musculus*	54
m169	*m169*	miR-27a and miR-27b	Murine cytomegalovirus	55, 56

for the prediction and validation of ceRNA interactions that is widely applicable to any potential transcript of interest.

One of the first examples linking aberrant ceRNA expression to tumorigenesis *in vivo* came from the discovery of a significant enrichment of predicted *PTEN* ceRNAs among genes whose loss accelerated melanomagenesis in a Sleeping Beauty insertional mutagenesis screen *in vivo*. *ZEB2* was subsequently validated as a *bona fide PTEN* ceRNA that modulated PTEN protein levels in an miRNA-dependent, protein-coding-independent fashion[18]. Decreased *ZEB2* transcript levels activated downstream signalling, enhanced cell transformation and were found to occur frequently in human melanoma and other cancers with low *PTEN* mRNA expression. These data suggest that the dysregulation of *PTEN* expression owing to the loss of its ceRNA *ZEB2* contributes to the development of melanoma both *in vitro* and *in vivo*.

Analysis of gene expression data in glioblastoma in combination with matched miRNA profiles validated 13 *PTEN* ceRNAs or miR program-mediated post-transcriptional regulatory (mPR) regulators whose locus deletions were predictive of decreased *PTEN* expression, downregulated *PTEN* in a 3′ UTR-dependent manner and increased tumour cell growth rates[26]. When the analysis was significantly extended beyond the binary ceRNA associations described in most other studies, the *PTEN* ceRNA interactions were found to be part of a post-transcriptional regulatory layer comprising more than 248,000 miRNA-mediated interactions.

The *VCAN* 3′ UTR has been reported to modulate PTEN levels by sequestering the shared miRNAs miR-144 and miR-136, freeing *PTEN* mRNA for translation[50]. In addition to its role as a *PTEN* ceRNA, *VCAN* has been shown to modulate the expression of several other genes through miRNA competition (Fig. 1). *VCAN* was also demonstrated to act as a ceRNA for the cell-cycle regulator RB1 by regulating miR-199a-3p and miR-144 levels, upregulating the expression of this crucial tumour suppressor both *in vitro* and *in vivo*[50]. *CD34* and *FN1* were validated as two additional *VCAN* ceRNAs that communicate through interactions with miR-133a, miR-199a-3p, miR-144 and miR-431 (refs 27, 51). As overexpression of the *VCAN* 3′ UTR *in vivo* was shown to induce organ adhesion followed by hepatocellular carcinoma development at a later time point, these studies were the first to demonstrate that the perturbation of a non-coding RNA transcript with validated ceRNA function has physiological and pathophysiological relevance *in vivo*.

Another gene with validated protein-coding ceRNAs is *CD44*, which encodes a transmembrane glycoprotein that is involved in a wide range of cellular functions[52]. The *CD44* 3′ UTR was shown to inhibit cell proliferation, colony formation, tumour growth and enhance cell motility, invasion and adhesion[52,53]. *CD44* regulated levels of CDC42, a Rho-GTPase involved in the control of cell morphology, migration and cell-cycle progression, by binding and sequestering three miRNAs, miR-216a, miR-330 and miR-608. It was also shown to modulate levels of Col1α1 through miR-328 binding and FN1 through miR-512-3p, miR-491 and miR-671 binding. The ceRNA interactions between *CD44* and *FN1* link *CD44*, *CDC42* and *Col1α1* to the broader miRNA–ceRNA network revolving around *PTEN* and *VCAN*. This adds another level of complexity to these post-transciptional regulatory associations between multiple transcripts, which play important parts in tumorigenesis (Fig. 1).

A recent study outlined the role of *Hmga2*'s ceRNA function in promoting lung carcinogenesis[54]. *Hmga2*, a non-histone chromosomal high mobility group protein, was shown to be highly expressed in metastatic lung adenocarcinoma and to promote lung cancer progression by sequestering the abundant *let-7* family of miRNAs. The TGF-β co-receptor *Tgfbr3* was identified as a putative *Hmga2* ceRNA, and *Tgfbr3*-driven TGF-β signalling was demonstrated to be largely necessary for *Hmga2*-mediated lung-cancer progression. The discovery that the primary function of a protein-coding transcript is to act as an oncogenic ceRNA for a very abundant miRNA, largely independently of its protein-coding function, provides further support for the hypothesis that the ceRNA function of multiple protein-coding transcripts is of fundamental importance in cancer progression.

A viral mRNA that functions as a natural miRNA sponge in host cells has also been described. Intriguingly, this mRNA was found to sequester the same miRNA reported to be downregulated by HSUR1-binding in *H. saimiri*, suggesting that ceRNA-mediated regulation of this miRNA may be a conserved mechanism for controlling host gene expression in multiple viruses. Two studies independently showed that the highly abundant murine cytomegalovirus mRNA *m169* acted as a natural miRNA sponge in host cells by binding to miR-27 through a single site in its 3′ UTR, hence directing its degradation[55,56]. Importantly, *m169*-mediated miR-27 regulation is another example of the importance of ceRNA function *in vivo*: disruption or replacement of the miRNA target site resulted in significant viral attenuation in multiple organs, suggesting that this miRNA-sponge interaction is crucial for efficient replication *in vivo*.

Predicting ceRNA crosstalk

ceRNA crosstalk depends on the MREs located on each transcript, which combinatorially form the foundation of these co-regulatory interactions[11]. Prediction of ceRNA crosstalk is thus dependent on the identification of MREs on the relevant transcripts of interest. Several miRNA-target prediction algorithms, including TargetScan, miRanda, rna22 and PITA, have been successfully used to identify ceRNA interactions. However, miRNA target identification is challenging owing to the imperfect nature of base pairing between an miRNA and its target, and the rules of targeting are not completely understood[57]. Further development and optimization of these algorithms will undoubtedly improve subsequent predictions of ceRNA interactions.

A database of predicted ceRNA interactions (ceRDB), which was generated by examining the co-occurrence of MREs on a genome-wide basis[58], has recently been compiled. Predicted miRNA–mRNA target interactions were obtained from TargetScan release 5.2, and interaction scores were defined for each mRNA by adding the total number of MREs that overlap with the miRNAs for the mRNA of interest; these interaction scores were then used to rank the predicted potential ceRNAs. Although this analysis is limited to the 3′ UTRs of protein-coding transcripts, it still represents a useful tool for the identification of putative ceRNA crosstalk.

As an alternative to *in silico* prediction strategies, recently developed high-throughput biochemical techniques, which identify endogenous miRNA–target interactions can be used. Examples of these experimental methods include high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP)[59] and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP)[60]. HITS-CLIP analysis of argonaute-bound miRNA–mRNA complexes has generated genome-wide interaction maps for specific miRNAs[61]. Recently, integration of HITS-CLIP data was shown to enhance miRNA target predictions by more than 20-fold over computational approaches alone, and mRNA–miRNA predictions have been used to identify candidate ceRNA interactions in the regenerating liver[62]. PAR-CLIP is a modification of the HITS-CLIP methodology with improved RNA recovery, which is able to indicate the exact targeting site more precisely. A complementary approach to these is MS2-tagged RNA affinity purification (MS2-TRAP), which can be used to identify all miRNAs associated with a target transcript in a particular cellular context[63]. Harnessing these experimental techniques will provide further insight into ceRNA regulation beyond that which is possible with *in silico* target predictions.

Additional considerations for ceRNA activity

Although multiple examples of ceRNA interactions have been described, little is known about the molecular conditions necessary for optimal ceRNA activity. Considerations such as the abundance of key players, potential interplay with RBPs and RNA editing may have profound effects on ceRNA crosstalk.

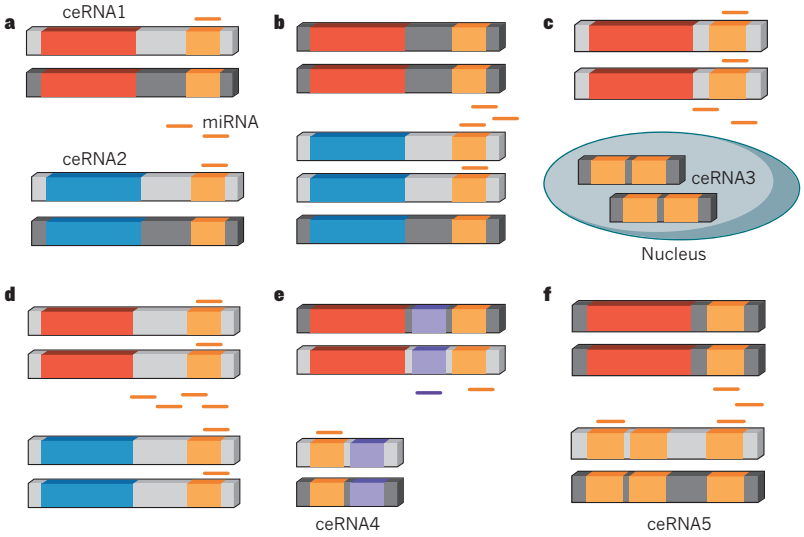


Figure 2 | Variable factors that may influence competing endogenous RNA (ceRNA) effectiveness. **a**, Steady state levels of ceRNA1 (protein-coding region shown in red) and ceRNA2 (protein-coding region shown in blue). Active repression by miRNAs is shown in light grey. **b**, If expression of ceRNA2 increases, this will increase expression of ceRNA1. **c**, ceRNA3 has a different subcellular localization from ceRNA1, and may be a less effective ceRNA than ceRNA2. **d**, Increased expression of the shared miRNA will increase repression of both ceRNA1 and ceRNA2. **e**, ceRNA4 contains miRNA response elements (MREs, purple and orange) for multiple shared miRNAs, and may be a more effective ceRNA than ceRNA2. **f**, ceRNA5 contains more MREs than ceRNA2 and may be a more effective ceRNA.

Abundance of miRNAs, ceRNAs and argonaute

Various factors, including miRNA- and ceRNA-expression levels and subcellular localization, the number of shared miRNAs and MREs, as well as binding affinity of the shared MREs have been suggested to contribute to ceRNA effectiveness[10,11,64] (Fig. 2). For example, it has been hypothesized that miRNA–RNA competition would apply only to a small subset of miRNAs whose abundance and corresponding target abundance fell within a narrow range, as the expression of competitor RNAs would have little impact on regulation by highly abundant miRNAs, and low abundance miRNAs would be unlikely to contribute much to gene regulation because a minimal number of targets would be bound at any given time[13].

Recent studies have further refined the dynamics and constraints of ceRNA crosstalk. Optimal conditions for ceRNA activity *in silico* have been determined using a mathematical mass-action model, and these were confirmed experimentally using *PTEN* and its validated ceRNA *VAPA*[65]. In addition, a minimal rate equation-based model to describe crosstalk between ceRNAs at steady state has been developed[66]. The relative abundance of both ceRNAs and miRNAs, their stoichiometry, the number of shared MREs and indirect interactions were shown to be crucial determinants of ceRNA cross-regulation. For instance, both studies found that ceRNA crosstalk was optimal when the transcript abundance of miRNAs and ceRNAs within a network were near equimolarity. ceRNA crosstalk would be minimal when ceRNA transcript abundance vastly exceeded that of miRNAs owing to the limited number of miRNAs available to repress their targets, as well as when miRNA transcript abundance vastly exceeded that of ceRNAs owing to maximal repression of most target transcripts.

Strong correlations in network connectivity that enhance ceRNA crosstalk have also been observed[66]: ceRNA interactions may be mediated by a large number of miRNA species that individually have weak effects on their respective targets; and ceRNA regulation may be symmetrical (two ceRNAs reciprocally regulate each other) or asymmetrical (one ceRNA regulates another but not vice versa). Intriguingly, a separate study[65] found that ceRNA and transcription factor networks were intricately intertwined and, with optimal molecular conditions, changes in the levels of a single ceRNA could affect integrated ceRNA and transcriptional networks.

Another potential rate-limiting determinant of ceRNA crosstalk is the abundance of argonaute, the catalytic component of the RNA-induced silencing complex (RISC). A mathematical model constructed to analyse changes in gene expression caused by competition between small RNAs, showed that competition for argonaute binding occurs within an intermediate range of argonaute abundance, whereby lower amounts of argonaute result in stronger competition[67]. Surprisingly, this model also revealed that when argonaute is highly abundant, the overexpression of one miRNA decreases expression levels of its own targets as well as the targets of other miRNAs. This is a result of the overexpressed miRNA binding to targets that share MREs for that miRNA and other miRNAs, freeing the other miRNAs to regulate other target genes.

Interplay between miRNAs and RBPs

As already discussed, in addition to competing for shared miRNAs, RNA transcripts may also compete for shared RBPs. These two regulatory functions are not always distinct and may in fact be intricately intertwined. RBP and miRNA crosstalk may be either direct, whereby they compete or cooperate for binding to a specific binding site, or indirect, whereby RBP binding changes the RNA secondary structure and thus alters accessibility to miRNA-binding sites. For example, competition between the stabilizing RBP HuR and miRNAs usually enhances gene expression if the HuR–mRNA association is predominant, and cooperation between HuR and miRNAs typically reduces expression of the target mRNA[68]. HuR has been shown to compete with miR-122, miR-548c, miR-494, miR-16 and miR-331-3p for binding to the *CAT1*, *TOP2A*, *NCL*, *COX2* and *ERBB2* mRNAs, respectively[69–73]. Conversely, HuR has been demonstrated to recruit the let-7 RISC and miR-19 RISC to repress the expression of *c-Myc* and *RhoB* transcripts, respectively[74,75]. Other examples of miRNA and RBP crosstalk include a study that reported that the germ-cell-specific RBP Dnd1 bound mRNAs and blocked miRNA access to their target sites[76]; and another that demonstrated that competition between miR-4661 and TTP for binding to the ARE sequence in IL-10 mRNA protected IL-10 from TTP-mediated degradation[77]. The exact RBP species present in a specific biological setting and their relative abundance may thus have a profound impact on miRNA and ceRNA regulatory networks.

RNA editing

RNA editing refers to a post-transcriptional processing mechanism producing an RNA sequence that is different from its template DNA[78]. The most prevalent type of RNA editing in higher eukaryotes is the

hydrolytic deamination of adenosine to inosine in double-stranded RNA substrates (A to I RNA editing). As inosine has the same base-pairing properties as guanosine, it pairs preferentially with cytidine instead of uridine. This effectively alters the sequence and base-pairing properties of the edited RNA, and may affect ceRNA regulation in two ways. First, editing of miRNA transcripts may repress biogenesis, affect strand stability or even alter their target spectrum[79]; second, editing of target RNA transcripts may abolish or create new miRNA-binding sites. One study reported the tissue-specific A to I editing of miR-376 cluster transcripts, and identified one editing site located in the middle of the 'seed' region, which is crucial for miRNA–target binding. Subsequently, the predominantly expressed edited miR-376 isoforms were shown to silence a different set of target genes compared with unedited miR-376 (ref. 80). This phenomenon is probably not limited to miR-376. In a second study, it was suggested that up to 16% of all human primary miRNAs may be targeted by ADARs[81]. A to I RNA editing was initially thought to be restricted to the coding regions of a few genes[79], but recent genome-wide screening coupled with bioinformatic analyses has revealed numerous editing sites that are frequently located in non-coding repetitive RNA sequences[78]. As 85% of pre-mRNAs are predicted to be subject to A to I RNA editing, with most target sites located in introns and UTRs[82,83], this phenomenon may significantly increase the complexity of elucidating ceRNA interactions.

Implications for human disease

Many of the genes with known ceRNA interactors identified so far have been implicated in human disease. For example, *PTEN* is a potent tumour suppressor gene that is frequently disrupted in multiple cancers and governs multiple cellular processes, including survival, proliferation, energy metabolism and cellular architecture[84]; *HULC* is the most upregulated gene in HCC and has been shown to regulate several genes involved in liver cancer[25,85]; T cells that are infected with HSUR1-expressing *H. saimiri* cause aggressive leukaemias and lymphomas[14] and *linc-MD1* expression in primary muscle cells isolated from patients with Duchenne muscular dystrophy has been shown to result in partial mitigation of the correct timing of the differentiation program[24]. This suggests that ceRNA crosstalk is not only of fundamental importance in physiological conditions, but is also crucially relevant in various pathophysiological states.

In addition to the genomic amplifications, deletions, mutations and epigenetic modifications that dysregulate crucial gene function and result in disease pathogenesis, aberrant changes in ceRNA regulation may also contribute to disease initiation and progression. First, changes in the expression levels of ceRNA transcripts represent an additional mechanism for regulating the levels of key disease genes, such as tumour suppressors or oncogenes. ceRNA crosstalk may thus allow the functional characterization of transcripts with no other known function on the basis of their ceRNA activity. This is of particular importance because although many lncRNAs (for example, *HULC*) are expressed at low levels in steady state conditions, their expression is known to be dysregulated in various diseases[6]. Second, it has been shown that, compared with non-transformed cell lines, cancer cells *in vitro* often express mRNAs with shorter 3′ UTRs, which result from alternative cleavage and polyadenylation[86]. This phenomenon has also been reported in primary breast and lung cancer tumours[87]. Transcripts with shorter 3′ UTRs are not only subject to reduced miRNA regulation in *cis*, but will have diminished ability to act as ceRNAs through miRNA crosstalk in *trans*. Third, the altered expression of transcript variants of many genes through alternative splicing or differential use of initiation and/or stop codons has been linked to disease[11,88,89]. These variants may produce proteins with different structures and functions, but may also introduce or eliminate MREs, and thus affect ceRNA activity. Fourth, chromosomal translocations are frequent events in cancer that may contribute to tumorigenesis. These events result in the production of various fusion proteins, but could also potentially affect ceRNA regulation owing to misexpression of 3′ UTRs under the control of non-endogenous promoters[11]. Finally, SNPs (single-nucleotide polymorphisms) associated with various diseases may create, modify or destroy MREs. For example, SNPs in genes related to Alzheimer's disease may potentially dysregulate miRNA and ceRNA networks and thus contribute to aberrant gene expression in Alzheimer's disease[90]. These observations suggest that ceRNA network interactions may have a role in determining the effectiveness of RNA-directed therapies[91].

Future perspectives

As ceRNA research is still in its infancy, there are only a handful of validated examples of each existing class of ceRNAs, such as mRNAs, small non-coding RNAs, lncRNAs, pseudogenes and the recently described circRNAs (Tables 1 and 2). Although research should certainly focus on the identification of more ceRNAs in these categories to ascertain whether ceRNA crosstalk represents a widespread network of RNA regulation, another exciting avenue for future work is the discovery of new classes of ceRNAs. For example, argonaute proteins have been found to interact with rRNAs and tRNAs, raising the possibility that these abundant, stable, small RNA species may also act as ceRNAs[92]. Intriguingly, 3′ UTRs that are expressed independently of their associated protein-coding sequences have been described in humans, mice and flies, suggesting that they may represent another new class of ceRNAs[93].

Refinements in ceRNA prediction strategies will facilitate the discovery of additional ceRNA interactions. One limitation of current ceRNA prediction strategies is that they do not factor in miRNA and potential ceRNA expression levels. This is especially crucial as miRNAs and ceRNAs are known to have temporal, spatial and disease-specific expression patterns, and it has been suggested that ceRNA crosstalk applies mainly to a subset of ceRNAs and miRNAs whose cellular concentration and target abundance fall within a specific range of values[10,11,13]. A precise understanding of miRNA and ceRNA abundance is thus essential for deconvoluting any ceRNA network of interest in a particular context. Although paired miRNA and mRNA expression profiles have been integrated with bioinformatic target predictions to refine potential ceRNA interactions[17,18,26], they have been based mainly on relative quantitation methods such as microarray and quantitative PCR (polymerase chain reaction) analysis that are subject to variations in probe binding and affinity. Techniques such as RNA sequencing that allow the quantitation of absolute levels of both ceRNA and miRNA transcript abundance will allow a more precise determination of ceRNA effectiveness. In addition, the integration of methods such as HITS-CLIP and PAR-CLIP, which allow the biochemical identification of endogenous miRNA–target interactions on a transcriptome-wide level in specific cellular contexts, will further refine ceRNA prediction strategies that currently depend mainly on *in silico* bioinformatic miRNA predictions.

Another limitation of current ceRNA prediction methods is their focus on crosstalk between canonical MREs that are located exclusively on the 3′ UTRs of target transcripts. A number of recent studies have suggested that miRNA regulation is not limited to canonical 3′ UTR targets. For example, several groups have demonstrated that a significant proportion of MREs in humans are non-canonical[94,95], and others have shown that effective miRNA targeting can occur outside the 3′ UTR[96,97]. Furthermore, it has been shown that miRNA binding does not always result in target downregulation[98]; 5′ UTR structure has been demonstrated to be a crucial determinant of miRNA-dependent regulation[99]; and, for dynamically recruited miRNAs, changes in overall miRNA expression were found to not always be correlated with miRNA recruitment to the RISC[62]. These new insights into miRNA function should be incorporated into future ceRNA prediction analysis, and may potentially add further complexities to the dynamics of ceRNA regulation and significantly expand the realm of potential ceRNA activity.

Although most of the ceRNA interactions identified so far have been between binary partners, there is increasing evidence that ceRNAs crosstalk in large interconnected networks; and that, in addition to direct interactions through shared miRNAs, secondary indirect

interactions may also have a profound effect on ceRNA regulation[26,65]. For example, *PTEN* expression has been shown to be regulated by ceRNA interactions between *PTEN* and its ceRNAs *PTENP1*, *CNOT6L*, *VAPA*, *VCAN* and *ZEB2* (Fig. 2). In addition, indirect ceRNA crosstalk (for example, between VCAN and FN1), as well as secondary interactions between *PTEN* ceRNAs (for example, between *PTENP1*, *CNOT6L* and *VAPA* mediated by the miR-17 and miR-19 families) may also contribute significantly to the effectiveness of ceRNA regulation. Future ceRNA research should extend beyond the identification of binary ceRNA interactions and include network analysis of potential intertwined miRNA and ceRNA networks.

The conservation of ceRNA crosstalk in multiple organisms, including plants, zebrafish, mice, humans and viruses, suggests that it may represent an important widespread layer of RNA regulation. As already discussed, this regulation is not limited to non-coding RNAs, it also encompasses protein-coding mRNAs. Recently, the genome of the carnivorous bladderwort *Utricularia gibba* was found to contain a typical number of protein-coding genes, but significantly less non-genic DNA[100], providing one example whereby a large non-coding genome is not essential for complex life. This presents an interesting opportunity to study potential RNA regulatory networks that involve mainly mRNAs and a much smaller non-coding transcriptome that consists of non-coding RNAs encoded in genic DNA such as circRNAs and antisense non-coding RNAs.

In summary, analysis of ceRNA interactions and crosstalk in intertwined networks may represent a robust platform to methodically functionalize non-coding transcripts on the basis of competition for shared miRNAs. Furthermore, the analysis sheds light on the non-coding function of protein-coding transcripts, and may uncover regulatory networks that have been overlooked by conventional protein-coding studies. To fully understand and appreciate the impact of ceRNA crosstalk on physiological and pathophysiological conditions, it will be of crucial importance to integrate these miRNA–RNA competitive networks with other competitive regulatory networks such as protein–RNA crosstalk. Although the full extent of ceRNA networks remains to be determined, this miRNA–RNA competition is a new post-transcriptional layer of endogenous competitive gene regulation, which has exciting implications for multiple basic biological systems, pathophysiological conditions and the development of new therapeutic approaches for cancer and other diseases. ■

Received 10 June; accepted 6 November 2013.

1. Volders, P. J. *et al*. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41,** D246–D251 (2013).
2. Dunham, I. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
3. Nagano, T. & Fraser, P. No-nonsense functions for long noncoding RNAs. *Cell* **145,** 178–181 (2011).
4. Brockdorff, N. *et al.* The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71,** 515–526 (1992).
5. Shabalina, S. A. & Spiridonov, N. A. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.* **5,** 105 (2004).
6. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* **9,** 703–719 (2012).
7. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2,** 986–991 (2001).
8. Mattick, J. S. & Gagen, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18,** 1611–1630 (2001).
9. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482,** 339–346 (2012).
10. Ebert, M. S. & Sharp, P. A. Emerging roles for natural microRNA sponges. *Curr. Biol.* **20,** R858–R861 (2010).
11. Salmena, L. *et al.* A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–358 (2011).
This essay proposes that all types of RNA transcripts may crosstalk and be co-regulated through a predictable 'ceRNA language' of RNA, and summarizes the experimental evidence in support of the ceRNA hypothesis at the time.
12. Seitz, H. Redefining microRNA targets. *Curr. Biol.* **19,** 870–873 (2009).
13. Wee, L. M., Flores-Jasso, C. F., Salomon, W. E. & Zamore, P. D. Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell* **151,** 1055–1067 (2012).
This work investigates the effect of miRNA and sponge transcript abundance on the efficacy of miRNA target repression and proposes that the ceRNA hypothesis may only apply to a subset of moderate or low abundance miRNAs.
14. Cazalla, D., Yario, T. & Steitz, J. A. Down-regulation of a host microRNA by a *Herpesvirus saimiri* noncoding RNA. *Science* **328,** 1563–1566 (2010).
This work extends ceRNA crosstalk to viruses and shows that a viral non-coding RNA sequesters and promotes the degradation of a miRNA in infected host cells.
15. Franco-Zorrilla, J. M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genet.* **39,** 1033–1037 (2007).
This study in plants, which demonstrates effective sequestration of miR-399 by the non-coding transcript IPS1, was the first to describe a role for non-coding RNAs as natural microRNA sponges in eukaryotic cells.
16. Poliseno, L. *et al*. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465,** 1033–1038 (2010).
This work focusing on the *PTEN* pseudogene was the first to define a functional role for expressed pseudogenes in cancer as well as attribute a ceRNA function to pseudogene and protein-coding mRNAs.
17. Tay, Y. *et al*. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* **147,** 344–357 (2011).
This study presents a combined computational and experimental approach that can be used as a framework to predict and validate ceRNA interactions.
18. Karreth, F. A. *et al. In vivo* identification of tumor-suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147,** 382–395 (2011).
This work used a Sleeping Beauty insertional mutagenesis screen to identify and validate ceRNAs *in vivo*.
19. Sanchez-Diaz, P. & Penalva, L. O. Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. *RNA Biol.* **3,** 101–109 (2006).
20. Lal, A. *et al*. Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *EMBO J.* **23,** 3092–3102 (2004).
21. Barker, A. *et al*. Sequence requirements for RNA binding by HuR and AUF1. *J. Biochem.* **151,** 423–437 (2012).
22. Al-Ahmadi, W. *et al*. Alternative polyadenylation variants of the RNA binding protein, HuR: abundance, role of AU-rich elements and auto-regulation. *Nucleic Acids Res.* **37,** 3612–3624 (2009).
23. Yoo, S. *et al*. A HuD-ZBP1 ribonucleoprotein complex localizes GAP-43 mRNA into axons through its 3′ untranslated region AU-rich regulatory element. *J. Neurochem.* **126,** 792–804 (2013).
24. Cesana, M. *et al*. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147,** 358–369 (2011).
25. Wang, J. *et al*. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res.* **38,** 5366–5383 (2010).
26. Sumazin, P. *et al*. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147,** 370–381 (2011).
This work in glioblastoma demonstrates that ceRNA interactions between approximately 7,000 transcripts are part of a surprisingly extensive post-transcriptional regulatory layer.
27. Lee, D. Y. *et al*. A 3′-untranslated region (3′UTR) induces organ adhesion by regulating miR-199a* functions. *PLoS ONE* **4,** e4527 (2009).
28. Brown, B. D. *et al*. Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nature Biotechnol.* **25,** 1457–1467 (2007).
29. Carè, A. *et al*. MicroRNA-133 controls cardiac hypertrophy. *Nature Med.* **13,** 613–618 (2007).
30. Ebert, M. S., Neilson, J. R. & Sharp, P. A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature Methods* **4,** 721–726 (2007).
This study is the first to describe the use of an experimental miRNA sponge to inhibit miRNA function.
31. Ebert, M. S. & Sharp, P. A. MicroRNA sponges: progress and possibilities. *RNA* **16,** 2043–2050 (2010).
32. Kluiver, J. *et al.* Generation of miRNA sponge constructs. *Methods* **58,** 113–117 (2012).
33. Gentner, B. *et al*. Stable knockdown of microRNA *in vivo* by lentiviral vectors. *Nature Methods* **6,** 63–66 (2009).
34. Valastyan, S. *et al*. A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis. *Cell* **137,** 1032–1046 (2009).
35. Brown, B. D. & Naldini, L. Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. *Nature Rev. Genet.* **10,** 578–585 (2009).
36. Haraguchi, T., Ozaki, Y. & Iba, H. Vectors expressing efficient RNA decoys achieve the long-term suppression of specific microRNA activity in mammalian cells. *Nucleic Acids Res.* **37,** e43 (2009).
37. Johnsson, P. *et al*. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature Struct. Mol. Biol.* **20,** 440–446 (2013).
38. Mitrovich, Q. M. & Anderson, P. mRNA surveillance of expressed pseudogenes in *C. elegans*. *Curr. Biol.* **15,** 963–967 (2005).
39. Marques, A. C. *et al*. Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. *Genome Biol.* **13,** R102 (2012).

40. Yoon, J. H. *et al*. LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* **47,** 648–655 (2012).
41. Fan, M. *et al*. A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. *Exp. Ther. Med*. **5,** 1143–1146 (2013).
42. Kallen, A. N. *et al*. The imprinted H19 lncRNA antagonizes Let-7 microRNAs. *Mol. Cell* **52,** 101–112 (2013).
43. Wang, Y. *et al*. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* **25,** 69–80 (2013).
44. Capel, B. *et al.* Circular transcripts of the testis-determining gene *Sry* in adult mouse testis. *Cell* **73,** 1019–1030 (1993).
45. Hansen, T. B. *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* **495,** 384–388 (2013).
This study was one of the first to functionalize circRNAs as a new class of highly stable and efficient natural miRNA ceRNAs.
46. Memczak, S. *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495,** 333–338 (2013).
47. Jeck, W. R. *et al*. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19,** 141–157 (2013).
48. Taulli, R., Loretelli, C. & Pandolfi, P. P. From pseudo-ceRNAs to circ-ceRNAs: a tale of cross-talk and competition. *Nature Struct. Mol. Biol.* **20,** 541–543 (2013).
49. Wilusz, J. E. & Sharp, P. A. Molecular biology. A circuitous route to noncoding RNA. *Science* **340,** 440–441 (2013).
50. Lee, D. Y. *et al.* Expression of versican 3′-untranslated region modulates endogenous microRNA functions. *PLoS ONE* **5,** e13599 (2010).
51. Fang, L. *et al*. Versican 3′-untranslated region (3′-UTR) functions as a ceRNA in inducing the development of hepatocellular carcinoma by regulating miRNA activity. *FASEB J*. **27,** 907–919 (2013).
52. Jeyapalan, Z. *et al.* Expression of CD44 3′-untranslated region regulates endogenous microRNA functions in tumorigenesis and angiogenesis. *Nucleic Acids Res*. **39,** 3026–3041 (2011).
53. Rutnam, Z. J. & Yang, B. B. The non-coding 3′ UTR of CD44 induces metastasis by regulating extracellular matrix functions. *J. Cell Sci.* **125,** 2075–2085 (2012).
54. Kumar, M. S. *et al.* Hmga2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature* http://dx.doi.org/10.1038/nature12785 (2013).
This study was the first to demonstrate that an abundant mRNA protein-coding transcript can act as ceRNA to out compete an abundant miRNA such as Let7.
55. Libri, V. *et al*. Murine cytomegalovirus encodes a miR-27 inhibitor disguised as a target. *Proc. Natl Acad. Sci. USA* **109,** 279–284 (2012).
56. Marcinowski, L. *et al*. Degradation of cellular mir-27 by a novel, highly abundant viral transcript is important for efficient virus replication *in vivo*. *PLoS Pathog*. **8,** e1002510 (2012).
57. Thomas, M., Lieberman, J. & Lal, A. Desperately seeking microRNA targets. *Nature Struct. Mol. Biol*. **17,** 1169–1174 (2010).
58. Sarver, A. L. & Subramanian, S. Competing endogenous RNA database. *Bioinformation* **8,** 731–733 (2012).
59. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460,** 479–486 (2009).
60. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–141 (2010).
61. Thomson, D. W., Bracken, C. P. & Goodall, G. J. Experimental strategies for microRNA target identification. *Nucleic Acids Res*. **39,** 6845–6853 (2011).
62. Schug, J. *et al.* Dynamic recruitment of microRNAs to their mRNA targets in the regenerating liver. *BMC Genomics* **14,** 264 (2013).
63. Yoon, J. H., Srikantan, S. & Gorospe, M. MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. *Methods* **58,** 81–87 (2012).
64. Mukherji, S. *et al.* MicroRNAs can generate thresholds in target gene expression. *Nature Genet.* **43,** 854–859 (2011).
65. Ala, U. *et al.* Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc. Natl Acad. Sci. USA* **110,** 7154–7159 (2013).
66. Figliuzzi, M., Marinari, E. & De Martino, A. MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory. *Biophys. J.* **104,** 1203–1213 (2013).
67. Loinger, A. *et al*. Competition between small RNAs: a quantitative view. *Biophys. J.* **102,** 1712–1721 (2012).
68. Srikantan, S., Tominaga, K. & Gorospe, M. Functional interplay between RNA-binding protein HuR and microRNAs. *Curr. Protein Pept. Sci.* **13,** 372–379 (2012).
69. Bhattacharyya, S. N. *et al*. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell* **125,** 1111–1124 (2006).
70. Srikantan, S. *et al*. Translational control of TOP2A influences doxorubicin efficacy. *Mol. Cell. Biol*. **31,** 3790–3801 (2011).
71. Tominaga, K. *et al*. Competitive regulation of nucleolin expression by HuR and miR-494. *Mol. Cell. Biol.* **31,** 4219–4231 (2011).
72. Young, L. E. *et al*. The mRNA stability factor HuR inhibits microRNA-16 targeting of COX-2. *Mol. Cancer Res.* **10,** 167–180 (2012).
73. Epis, M. R. *et al*. The RNA-binding protein HuR opposes the repression of *ERBB-2* gene expression by microRNA miR-331–3p in prostate cancer cells. *J. Biol. Chem*. **286,** 41442–41454 (2011).
74. Kim, H. H. *et al.* HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev*. **23,** 1743–1748 (2009).
75. Glorian, V. *et al*. HuR-dependent loading of miRNA RISC to the mRNA encoding the Ras-related small GTPase RhoB controls its translation during UV-induced apoptosis. *Cell Death Differ.* **18,** 1692–1701 (2011).
76. Kedde, M. *et al*. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell* **131,** 1273–1286 (2007).
77. Ma, F. *et al*. MicroRNA-466l upregulates IL-10 expression in TLR-triggered macrophages by antagonizing RNA-binding protein tristetraprolin-mediated IL-10 mRNA degradation. *J. Immunol*. **184,** 6053–6059 (2010).
78. Nishikura, K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nature Rev. Mol. Cell Biol.* **7,** 919–931 (2006).
79. Maas, S. Posttranscriptional recoding by RNA editing. *Adv. Protein Chem. Struct. Biol.* **86,** 193–224 (2012).
80. Kawahara, Y. *et al.* Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315,** 1137–1140 (2007).
81. Kawahara, Y. *et al.* Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res*. **36,** 5270–5280 (2008).
82. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol*. **2,** e391 (2004).
83. Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature Biotechnol.* **22,** 1001–1005 (2004).
84. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nature Rev. Mol. Cell Biol.* **13,** 283–296 (2012).
85. Panzitt, K. *et al*. Characterization of *HULC*, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology* **132,** 330–342 (2007).
86. Mayr, C. & Bartel, D. P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138,** 673–684 (2009).
87. Lembo, A., Di Cunto, F. & Provero, P. Shortening of 3′UTRs correlates with poor prognosis in breast and lung cancer. *PLoS ONE* **7,** e31129 (2012).
88. Pal, S., Gupta, R. & Davuluri, R. V. Alternative transcription and alternative splicing in cancer. *Pharmacol. Ther.* **136,** 283–294 (2012).
89. Venables, J. P. *et al.* Cancer-associated regulation of alternative splicing. *Nature Struct. Mol. Biol.* **16,** 670–676 (2009).
90. Mallick, B. & Ghosh, Z. A complex crosstalk between polymorphic microRNA target sites and AD prognosis. *RNA Biol*. **8,** 665–673 (2011).
91. Almeida, M. I., Reis, R. M. & Calin, G. A. Decoy activity through microRNAs: the therapeutic implications. *Expert Opin. Biol. Ther.* **12,** 1153– 1159 (2012).
92. Tollervey, D. Molecular biology: RNA lost in translation. *Nature* **440,** 425–426 (2006).
93. Mercer, T. R. *et al*. Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res.* **39,** 2393–2403 (2011).
94. Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature Methods* **10,** 253–255 (2013).
95. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153,** 654–665 (2013).
96. Tay, Y. *et al.* MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* **455,** 1124–1128 (2008).
97. Hausser, J., Syed, A. P., Bilen, B. & Zavolan, M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* **23,** 604–615 (2013).
98. Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318,** 1931–1934 (2007).
99. Meijer, H. A. *et al.* Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation. *Science* **340,** 82–85 (2013).
100. Ibarra-Laclette, E. *et al*. Architecture and evolution of a minute plant genome. *Nature* **498,** 94–98 (2013).

Acknowledgements We thank S. M. Tan, R. Taulli, F. Karreth and other members of the Pandolfi laboratory for helpful discussions and critical review of the manuscript. Y.T. received a Special Fellow Award from The Leukemia & Lymphoma Society. P.P.P. was supported by US National Institutes of Health grant R01 CA-82328.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at go.nature.com/emoqic. Correspondence should be addressed to P.P.P. (ppandolf@bidmc.harvard.edu).

REVIEW

doi:10.1038/nature12987

PIWI proteins and PIWI-interacting RNAs in the soma

Robert J. Ross[1]*, Molly M. Weiner[1]* & Haifan Lin[1]

The discovery of millions of PIWI-interacting RNAs revealed a fascinating and unanticipated dimension of biology. The PIWI–piRNA pathway has been commonly perceived as germline-specific, even though the somatic function of PIWI proteins was documented when they were first discovered. Recent studies have begun to re-explore this pathway in somatic cells in diverse organisms, particularly lower eukaryotes. These studies have illustrated the multifaceted somatic functions of the pathway not only in transposon silencing but also in genome rearrangement and epigenetic programming, with biological roles in stem-cell function, whole-body regeneration, memory and possibly cancer.

The discoveries of small non-coding RNAs, including PIWI-interacting RNAs (piRNAs), have significantly expanded the RNA world. piRNAs are generally 24–32 nucleotides in length and bind specifically to the PIWI subfamily of Argonaute proteins. Piwi was originally discovered in *Drosophila*[1], in which it functions in germline stem-cell maintenance and self-renewal[2]. For clarity, we use PIWIs to refer collectively to PIWI proteins, whereas Piwi refers to the individual protein. Although piRNAs were discovered and formally defined in mammalian systems in 2006 as small non-coding RNAs that specifically interact with PIWIs[3–6], cloning of piRNAs in *Drosophila*[7–9] revealed that they include a previously discovered class of small non-coding RNAs called repeat-associated RNAs (rasiRNAs)[10,11]. Since 2006, the PIWI–piRNA field has rapidly advanced, with a focus on the germ line, in which PIWIs and piRNAs are enriched and PIWI mutations lead to a profound infertility phenotype[12–14]. Indeed, the name PIWI comes from the original mutant phenotype P-element-induced wimpy testis[1]. The best-known role of the PIWI–piRNA pathway in the germ line is in transposon silencing, because piRNAs map largely to transposable elements[9], with PIWI depletion leading to a drastic increase in transposon messenger RNA expression[15].

Despite the germline focus, since their discovery, PIWIs' somatic function has long been documented. Initial work on the *Drosophila* gene *piwi*, the first identified member that defines the *Argonaute* gene family, determined that its germline function depends on the somatic cells of the gonad[2]. Recently, significant insight into the somatic function of the PIWI–piRNA pathway has come from the ovarian somatic cells of *Drosophila*. In addition, groundbreaking work in lower eukaryotes has demonstrated a conserved function for PIWIs and their associated piRNAs in somatic tissues — particularly in stem cells. In this Review, we focus on the role of the PIWI–piRNA pathway in the soma of diverse organisms, from basal eukaryotes to humans. We begin by looking at PIWI expression and piRNA biogenesis in somatic tissues, and then illustrate how the PIWI–piRNA pathway exerts diverse functions, including epigenetic regulation, transposon silencing, genome rearrangement and developmental regulation. Through this Review we hope to illustrate a broader role for the PIWI–piRNA pathway in the soma.

Expression of PIWIs and piRNAs in somatic tissues

PIWIs are expressed in organisms from sponges to humans. This expression occurs in remarkably diverse cell types, ranging from naive pluripotent stem cells to differentiated somatic cells, with most somatic expression related to various totipotent and pluripotent stem cells[16–24] (Table 1). Several studies have demonstrated an essential stem-cell function for PIWIs. One interesting example comes from planarian stem cells. *piwi* genes are expressed in planarian totipotent stem cells, called neoblasts, that can repopulate all somatic and germline lineages. Planarians have an unusual regenerative capability, in which neoblasts are responsible for maintaining and regenerating all tissues. PIWIs play a crucial part in the neoblast lineage, as illustrated by the two phenotypic consequences of knockdown of mRNAs encoding PIWIs: first, animals are incapable of body-part regeneration; and second, failure of tissue maintenance ultimately results in death[21]. Furthermore, piRNA-like small RNAs, which depend on PIWIs for their biogenesis, have been identified in planaria. These RNAs are not germline-restricted, because ablation of the germ line by *nanos* mRNA knockdown does not affect piRNA-like RNA production[21,22]. In addition, RNA interference (RNAi) experiments in ascidians have demonstrated the requirement for PIWIs in whole-body regeneration (discussed later)[23,24].

Although substantial work has described PIWI expression beyond the germ line of lower eukaryotes, the somatic function of PIWIs was first characterized in somatic cells of the *Drosophila* ovary. The *Drosophila* Piwi protein is expressed in all somatic cells within the ovary and in early somatic cells of the testis[2,25]. Outside the gonad, Piwi binds to polytene chromosomes of the salivary gland[26]. Emerging and unexpected evidence shows that PIWIs function outside the gonad, particularly in the *Drosophila* head. The first genetic evidence showed that *piwi* mutation leads to position effect variegation in the expression of the eye colour gene *white* in a clonal fashion[8,26]. This suggests the possibility of cell-autonomous Piwi function in the eye. Direct molecular evidence came from a more recent study in which two *Drosophila* PIWIs, Argonaute-3 and Aubergine, which were once thought to be germline-specific, showed region-specific expression in the brain, with their mutation leading to transposon upregulation in fly heads[27]. Interestingly, piRNA-like small RNAs mapping to transposons and heterochromatin are detectable in the *Drosophila* head on depletion of the RNAi effector Argonaute-2 (ref. 28). A recent publication has shown that all *Drosophila* PIWIs are expressed during early embryogenesis and that depletion of maternal *Drosophila* PIWIs results in profound chromosomal and mitotic defects, thus establishing a crucial function for PIWIs in the earliest stage of somatic development[29]. Together with the work on basal eukaryotes, studies in *Drosophila* highlight the existence and function of PIWIs in somatic tissues. Although somatic PIWIs

[1]Yale Stem Cell Center and Department of Cell Biology, Yale University School of Medicine, New Haven, Connecticut 06509, USA. *These authors contributed equally to this work.

Table 1 | Piwi orthologue expression

Phylum	Common name	Species	Known Piwi genes	PIWI protein expression	Reference
Porifera	Sponge	*Ephydatia fluviatilis*	*EfPiwiA* and *EfPiwiB*	Archeocytes (stem cells that differentiate into both somatic and germ cells)	16
Cnidaria	Jellyfish	*Clytia hemisphaerica*	*Piwi*	Somatic stem cells of the tentacle bulb (produce stinging cells characteristic of the cnidarians)	17
	Jellyfish	*Podocoryne carnea*	*Cniwi*	Somatic stem cells of the tentacle bulb (see above); striated muscle cells capable of transdifferentiation	18
Ctenophora	Comb jellyfish	*Pleurobrachia pileus*	*PpiPiwi1* and *PpiPiwi2*	Actively dividing adult somatic cells; germ line	19
Platyhelminthes	Planaria	*Schmidtea mediterranea*	*smedwi-1, smedwi-2* and *smedwi-3*	Neoblasts (totipotent stem cells that can repopulate all somatic and germline lineages)	21, 22
	Saltwater flatworm	*Macrostomum lignano*	*macpiwi*	Neoblasts (see above)	20
Mollusca	Sea slug	*Aplysia californica*	*Piwi*	Nervous system, heart and germ line	80
Arthropoda	Fruitfly	*Drosophila melanogaster*	*piwi, aub* and *AGO3*	Gonad, brain, salivary gland	2, 26, 92, 93
Chordata	Sea squirt (ascidian)	*Botrylloides leachii* and *Botryllus schlosseri*	*Piwi*	Stem cell population (capable of whole-body regeneration)	23, 24
	Mouse	*Mus musculus*	*Miwi*, *Mili* and *Miwi2*	Diverse cancers (breast cancer, rhabdomyosarcoma, medulloblastoma); male germ line	12–14, 33
	Human	*Homo sapiens*	*HIWI, HILI, HIWI2* and *HIWI3*	Diverse cancers (breast cancer, cervical cancer, endometrial carcinoma, seminomas, hepatocellular carcinoma, gastric cancer, pancreatic adenocarcinoma, gastrointestinal stromal tumours, colon cancer, renal cell carcinoma); haematopoietic stem cells; and male germ line	30, 32–36

clearly function in *Drosophila*, it remains a mystery whether piRNAs are generated outside the gonad, because some of the most studied piRNA biogenesis factors are not robustly expressed in non-gonadal somatic tissue (discussed later).

In light of the above findings, an important question remains: are PIWIs present in mammalian somatic tissues? There are four human PIWIs: HIWI (also known as PIWIL1), HILI (also known as PIWIL2), HIWI2 (also known as PIWIL4) and HIWI3 (also known as PIWIL3). HIWI is expressed in haematopoietic stem cells but not in their differentiated progeny. Although this once led to the suggestion that HIWI might be involved in the stemness of these stem cells[30], a powerful genetic experiment in which all three mouse *Piwi* genes were knocked out showed no detectable effect on haematopoiesis. Thus, PIWI expression in haematopoietic stem cells may not have any functional implication or PIWI function is redundant with other proteins, possibly Argonaute subfamily proteins[31]. With regards to human health, many studies demonstrate PIWI expression in a wide variety of human cancers; however, these data are at best correlative and it is too early to tell whether PIWIs have any role in cancer (discussed later)[32–36].

The expression of PIWIs in mammalian somatic tissues implies the potential existence of somatic piRNAs. One study provided evidence of piRNA expression in diverse somatic tissues of both the mouse and macaque[37]. However, the lack of appropriate negative controls combined with a small library size make it difficult to differentiate between true piRNA expression and contamination during library preparation. Indeed, this study might be a cautionary tale describing the challenges in the search for mammalian somatic piRNAs. Further work will help to determine whether piRNAs are expressed outside the germ line in mammals. Perhaps malignant tissues are good sites to begin the search given the expression of PIWIs in cancer.

piRNA biogenesis in the *Drosophila* ovarian soma

Our mechanistic understanding of somatic piRNA biogenesis comes from recent work in the somatic follicle cells of the *Drosophila* ovary. piRNAs in the ovarian soma are generated by a Piwi-dependent mechanism through the primary piRNA biogenesis pathway[38,39]. This process is independent of the other two *Drosophila* PIWIs, Aubergine and Argonaute-3, which function as a piRNA amplification loop in the germline cells (known as secondary piRNA biogenesis). In the primary piRNA biogenesis pathway, long piRNA precursors are transcribed from specific genomic loci known as piRNA clusters, cleaved and modified in the cytoplasm, and then transported into the nucleus in complex with Piwi (Fig. 1). Ovarian soma-specific piRNAs are transcribed from two main loci. The *flamenco* locus contains transposon remnants and encodes a long single-stranded piRNA precursor that is antisense to *flamenco*'s component transposons. Precursors generated from the *flamenco* locus target transposons of the gypsy family of long terminal repeat (LTR) transposons, including *gypsy*, *ZAM* and *idefix*[38]. Another ovarian soma-specific locus, *traffic jam*, has two functions: the 3′ UTR of its transcript is a substrate for somatic piRNA production, whereas its protein product drives Piwi expression in ovarian somatic cells[39].

It is unclear how long single-stranded piRNA precursors are exported from the nucleus and how these precursors are initially processed into smaller fragments in the cytoplasm. One proposed, but unproven, candidate for piRNA 5′ end determination is Zucchini, an outer mitochondrial membrane protein with single-strand-specific endonuclease activity *in vitro*[40]. In *Drosophila*, the maturing precursor then enters the perinuclear Yb body[41–43], an ovarian soma-specific cytoplasmic structure named after the Yb protein[44]. Although it is not yet clear what occurs in the Yb body, several of its components are crucial for the processing of piRNA precursors into mature piRNAs, and for subsequent Piwi nuclear localization. These components include the helicase Armitage[41–43] and the TUDOR-domain-containing protein Vreteno[44]. Both Vreteno[45] and Yb[41,43] are needed for Piwi expression and/or stability. Zucchini also functions in piRNA maturation. Knockdown of the mRNA encoding Zucchini in a *Drosophila* ovarian somatic cell line leads to cytoplasmic accumulation of piRNA-intermediate-like molecules[46] and accumulation of Piwi in the Yb body; correspondingly, nuclear Piwi is absent[42]. The mouse homologues of the above piRNA biogenesis factors, including Zucchini[47], Armitage[48,49], Shutdown[50,51] and Vreteno[52] are crucial for piRNA biogenesis in the testes, and male mutants are infertile, thus implying conservation of these factors in piRNA biogenesis.

Next, piRNAs are loaded onto Piwi in the cytoplasm through an uncharacterized step that is independent of Piwi slicer activity and nuclear localization signal[39,53]. The co-chaperone Shutdown, which is essential to piRNA biogenesis in both the ovarian soma and germ line, may function in piRNA loading[54]. Ovarian somatic piRNAs contain a characteristic 5′ U, but the mechanism generating this 5′ U bias is unknown. It is possible that Piwi itself preferentially binds RNAs with 5′ U, as is the case for silkworm Piwi *in vitro*[55]. This might authenticate

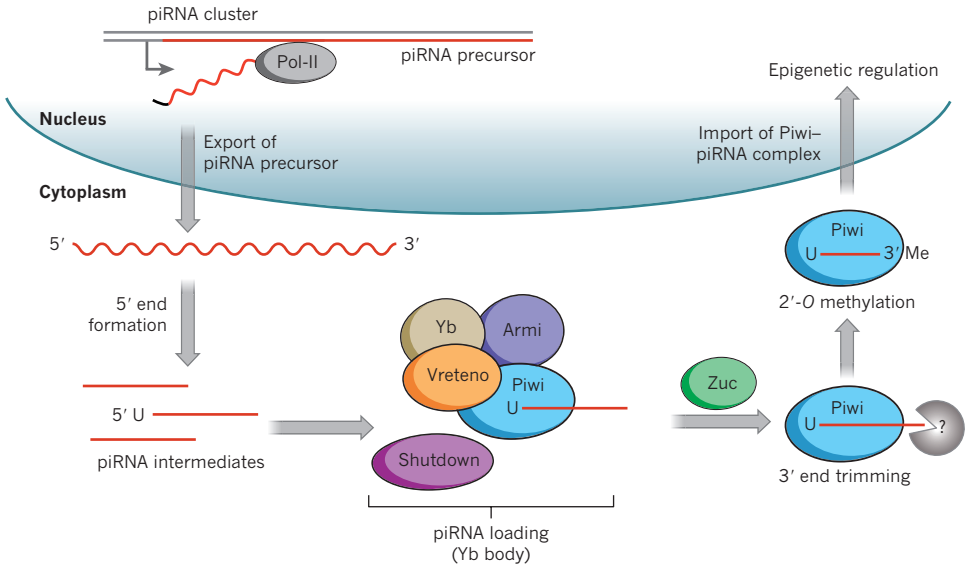


Figure 1 | piRNA biogenesis in the *Drosophila* ovarian soma. piRNAs are generated from specific genomic loci known as piRNA clusters, which include *flamenco*, the 5′ UTRs of mRNAs and *traffic jam* in the soma. The long single-stranded piRNA precursor (red) is then exported from the nucleus. In the cytoplasm, the precursors are processed into mature piRNAs. The precursors are cleaved by an unknown endonuclease to generate their 5′ end, and transported into the perinuclear Yb body for further processing. The precursors are loaded onto Piwi in a process that is dependent on Yb, Armitage (Armi) and Vreteno. Overlapping proteins indicate protein–protein interaction. The co-chaperone Shutdown plays an uncharacterized, but crucial, part in piRNA loading. The putative endonuclease Zucchini (Zuc) is required for piRNA maturation and for nuclear localization of Piwi. Subsequently, piRNAs are trimmed to the appropriate length by an unidentified exonuclease and 2′-*O*-methylated at the piRNA 3′ end, rendering them more stable. The Piwi–piRNA complex is then transported into the nucleus, where it modulates chromatin state.

and stabilize piRNA intermediates with 5′ U for further processing. An unknown protein then trims the 3′ ends of the maturing piRNAs. The shared structure of Argonaute family proteins suggests the possibility that piRNA length may be determined by the number of bases protected within Piwi[56–58]. This also explains the characteristic size profile of piRNAs bound by each of the three *Drosophila* PIWIs[9]. Mature piRNAs are then 2′-*O*-methylated by a conserved methyltransferase, Pimet[59], perhaps to ensure their stability. Finally, in an uncharacterized step, the Piwi–piRNA complex is imported into the nucleus to exert its regulatory function.

Epigenetic regulation by the PIWI–piRNA pathway

Strong evidence indicates that Piwi and piRNAs play a crucial part in epigenetic regulation. Pioneering work has shown that Piwi functions in the transcriptional silencing of *Adh* transgene arrays, and small RNAs have been proposed to have a role in Piwi-mediated silencing[60]. Piwi is a predominantly nuclear protein[25] that localizes to salivary gland polytene chromosomes in an RNA-dependent manner[26], and co-localizes with known epigenetic modifiers such as the Polycomb group proteins[61]. There is general agreement that Piwi is required for appropriate histone methylation and transcriptional gene silencing, as loss of Piwi in both germ line and soma leads to a decrease in the repressive methylation mark on histone 3 lysine 9 (H3K9)[62–66], an increase in RNA Pol II occupancy[63–65] and an increase in nascent transcript[60,63,66]. Indeed, tissue-specific knockdown of Piwi in somatic ovarian follicle cells implicates transcriptional repression as the dominant mode of transposable element control in the soma[66].

What is the mechanism by which Piwi directs epigenetic modification? The answer to this lies in piRNA, which probably guides Piwi to specific target sequences in the genome by sequence complementarity[63–66]. Chromatin immunoprecipitation (ChIP) data from our laboratory suggest that the Piwi–piRNA complex binds its genomic target in euchromatin through a nascent transcript (often a long non-coding RNA), and in heterochromatin predominantly through a direct piRNA–DNA interaction[26,64]. According to these results, which await reproduction by other labs, the Piwi–piRNA complex then recruits epigenetic factors such as HP1a and the histone methyltransferase Su(var)3-9 to exert their function (Fig. 2). HP1a is a highly conserved chromatin factor whose *N*-terminal chromodomain binds trimethylated H3K9 whereas the *C*-terminal chromoshadow domain dimerizes with the same domain in another HP1a molecule and interacts with other proteins[64]. Piwi may directly recruit HP1a[26]. HP1a may then recruit one of its well-characterized interactors, Su(var)3-9, which is responsible for most H3K9 methylation in *Drosophila*[64]. Alternatively, others have proposed that Piwi may first recruit a histone methyltransferase such as Su(var)3-9 or Setdb1. The methyltransferase could establish H3K9 methylation and in turn recruit HP1a; this possibility is supported by work in fission yeast[67]. In addition to these core epigenetic factors, Asterix (also known as DmGTSF1), an upstream nuclear factor, is required for Piwi-directed H3K9 methylation[68–70]. A recently identified downstream effector of Piwi, Maelstrom, is not required for the establishment of H3K9 methylation but is required for the transcriptional silencing of transposable elements[63].

Several studies have revealed unanticipated epigenetic gene regulation carried out by Piwi. In *Drosophila simulans*, maternally deposited piRNAs against the retrotransposon *tirant* initiated H3K9-mediated transcriptional gene silencing of this element in the somatic tissues of developing embryos[71]. This study shows that maternal germline-derived piRNAs are required for somatic epigenetic programming. A recent study expanded on this idea by showing that both maternal and zygotic Piwi are required for establishment of heterochromatin in non-gonadal somatic cells of the early embryo[72]. Interestingly, Piwi-targeting of transposable elements for silencing means that those genes containing or in proximity to transposable element sequences may be piRNA pathway targets. The presence of a transposon or its remnants in an intron or in proximity to a gene correlates with significant transcriptional repression[63,69]. Surprisingly, Piwi can also act as an epigenetic activator. In *Drosophila*, Piwi establishes euchromatic features of chromosome 3R telomere-associated sequence (3R-TAS)[8], and whole-genome studies have shown that Piwi binding may enhance transcriptional activation marks in multiple regions[64]. In support of this unexpected epigenetic function, a recent independent study confirmed that *piwi* mutant flies have increased HP1a enrichment in regions, including 3R-TAS, in which Piwi is implicated as an epigenetic activator[72]. How can Piwi-binding lead

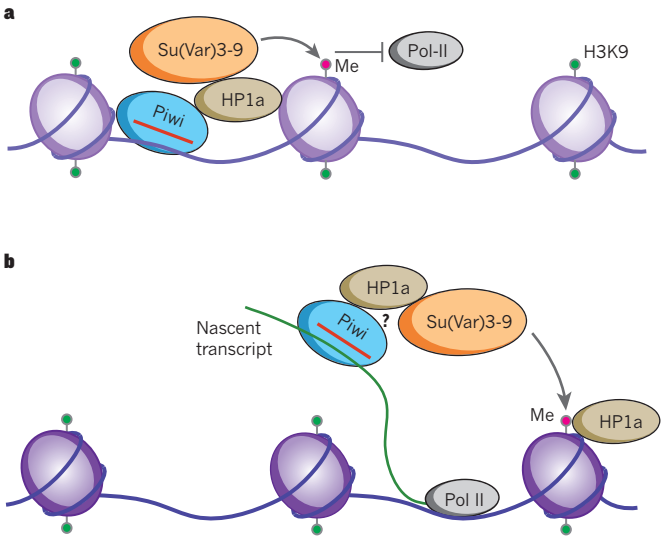


Figure 2 | Piwi–piRNA mediated epigenetic regulation. Simplified illustrations of the currently proposed models of Piwi-mediated transcriptional gene silencing. **a**, In heterochromatin, Piwi may be guided to its target sequences by the complementarity of its bound piRNA (red) to genomic DNA. On binding, Piwi recruits the epigenetic modifier HP1a that then recruits the major *Drosophila* histone methyltransferase Su(var)3-9, which then deposits a methyl group on the unmethylated histone 3 lysine 9 (H3K9). Through an unknown mechanism, a critical mass of the H3K9 repressive chromatin marks inhibits Pol II transcription, effectively silencing the Piwi–piRNA target. **b**, In euchromatin, Piwi targets nascent transcripts (green) by piRNA sequence complementarity. Subsequent to this, there are multiple working models. Piwi may directly recruit HP1a, which in turn recruits a histone methyltransferase (Su(var)3-9 is shown) that then deposits methyl groups on the unmethylated H3K9. Alternatively, Piwi may directly recruit the histone methyltransferase, and subsequently HP1a may then bind the methylated H3K9. Regardless of the mechanism, the net effect is that these repressive marks, in concert with Maelstrom (not shown), inhibit RNA Pol II transcription.

to transcriptional silencing in most target sites, but activation in a small number of target sites? Perhaps Piwi interacts with different partners and/or is influenced by the local chromatin micro-environment or its sublocalization within the nucleus. Clearly, Piwi has many different roles in a variety of cellular processes, and one challenge in the field is to unite or to further distinguish between these many functions.

PIWI–piRNA pathway in genome rearrangement

Ciliates are single-celled organisms that possess remarkable nuclear dimorphism. They have two distinct genomes: a somatic macronucleus that functions in vegetative growth and is actively transcribed, and a transcriptionally silent germline micronucleus that functions in the exchange of genetic information during sexual reproduction, known as conjugation. Following conjugation and mitosis, the zygotic genome in one of the two micronuclei in each daughter cell is extensively edited to create a new and partial somatic genome that replaces the parental germline genome, in a process known as genome rearrangement or, more appropriately, somatic elimination. During somatic elimination, repetitive sequences are removed from the genome before polyploidization, thus preventing their transcription in the resultant macronuclei[73]. Although other organisms have complex mechanisms for transcriptional and post-transcriptional silencing of transposable elements, the ciliates simply delete such sequences from the somatic macronuclear genome.

Early studies in the ciliate *Tetrahymena* revealed that their Piwi orthologues are essential for proper genome rearrangement and piRNA production. These piRNAs, termed scan RNAs in the original literature, are derived from the micronucleus during conjugation. Importantly, *Tetrahymena* piRNAs are distinct from those in other species as they are produced through a Dicer-dependent mechanism from double-stranded RNA precursors[74]. In this sense, they are more like endogenous small interfering RNAs[75] (endo-siRNAs). After mating, the piRNAs provide sequence-specificity to target germline-restricted regions in the developing somatic macronucleus for elimination (Fig. 3)[76]; consequent histone methylation then groups these sequences for collective deletion[77]. In the current model, for which there is strong evidence in both *Tetrahymena* and *Paramecium*, long non-coding RNAs transcribed from the parental somatic macronucleus act as sponges for germline-derived piRNAs. Thus, unbound piRNAs specifically mark any non-somatic regions in the genome for elimination in the developing daughter macronucleus[78].

By contrast, piRNAs in the ciliate *Oxytricha* complex with a Piwi orthologue, Otiwi1, to mark somatic genes for retention during development of the somatic macronucleus. PIWIs are so crucial to *Oxytricha* that on knockdown of the mRNA encoding Otiwi1 the organisms do not survive after mating. Furthermore, injection of RNAs that target normally deleted genes leads to their retention through multiple sexual generations[79]. This suggests that genomic composition changes of the somatic macronucleus are heritable across multiple generations. In *Tetrahymena*, the minority of the developing somatic genome is directed for deletion, whereas in *Oxytricha* the minority of the developing somatic genome is directed for retention. Although it is not known why these distantly related species solve this puzzle of genome rearrangement in opposite ways, it is clear that each has evolved the most efficient system with the fewest possible piRNA targets to direct the assembly of a complete somatic macronuclear genome[79].

PIWI–piRNA pathway in somatic development

The PIWI–piRNA pathway has broad developmental functions in diverse organisms, from memory in the sea slug *Aplysia* to whole-body regeneration in botryllids. In *Aplysia*, both synaptic plasticity and associative memory formation require the PIWI–piRNA pathway. *Aplysia* Piwi, in complex with piRNAs, responds to the neurotransmitter serotonin by directing CpG methylation of the CREB2 promoter. CREB2 is a major inhibitor of memory in *Aplysia*, so Piwi-mediated transcriptional silencing of CREB2 results in memory through long-lasting, cell-wide enhancement of synaptic transmission[80]. The capacity of the PIWI–piRNA pathway to epigenetically regulate genes is thus exerted in mature neurons to promote cellular memory, emphasizing PIWI function in differentiated somatic tissue.

Colonial ascidians are chordates that are capable of whole-body regeneration. Remarkably, the ascidian *Botrylloides leachii* can regenerate its entire body from any blood vessel fragment containing only a few cells. This phenomenon depends on a population of Piwi-positive cells present on the luminal side of the vascular epithelium. On RNAi knockdown of the mRNA encoding Piwi, organisms cannot undergo whole-body regeneration[23]. In the closely related *Botryllus schlosseri*, Piwi is also crucial for whole-body regeneration. Piwi positive-cells, which contribute to both the germline and somatic lineages of future generations, reside within the endostyle niche. *B. schlosseri* undergoes weekly asexual growth, during which a population of Piwi-positive stem cells vacates the original endostyle niche and migrates to the niche within the growing daughter organism before the parent niche undergoes massive apoptosis. In this way, the stem-cell population efficiently preserves itself over the course of the colony's lifetime through migration to the offspring[24].

Piwi suppresses phenotypic variation

The PIWI–piRNA pathway has a direct role in buffering against phenotypic variation, and Piwi depletion in *Drosophila* results in new somatic defects in a random fashion and at low frequency. Canalization, a term coined by Conrad Waddington in 1942 (ref. 81), describes developmental robustness: the ability of a system to generate a single phenotype regardless of genetic or environmental perturbations. The heat-shock protein Hsp90 is a crucial chaperone in the suppression of phenotypic variation, that is, in canalization. Depletion of Hsp90 generates diverse somatic phenotypic variants in species ranging from *Arabidopsis*[82] to *Drosophila*[83]. These variants are heritable even in the absence of heat

shock or other forms of stress, and even on repletion of the wild-type Hsp90 levels in progeny, indicating that the effect is not simply due to the chaperone function of Hsp90 during stress. *Drosophila* Piwi forms a complex with Hsp90 and the heat-shock organizing protein, Hop, *in vivo* to suppress phenotypic variation[84].

This buffering of variation is probably accomplished through both genetic and epigenetic mechanisms. Evidence for a genetic mechanism stems from a study in which Hsp90 depletion in *Drosophila* leads to defective transposable element silencing by the PIWI–piRNA pathway. The resultant transposon-mediated mutagenesis may generate new phenotypes[85]. However, the transposon mutagenesis hypothesis cannot explain the full spectrum of canalization phenotypes. Although one copy of maternal *piwi* is sufficient for transposon silencing, it is insufficient to suppress phenotypic variation[84]. In addition, maternal epigenetic factors, such as the *trithorax* group of genes that maintain active chromatin, also buffer against phenotypic variation in *Drosophila*[86]. These observations indicate the involvement of an epigenetic mechanism in suppressing phenotypic variation. With regards to the inheritance of such phenotypes, work in the *Caenorhabditis elegans* germ line has shown that piRNAs induce transgenerational epigenetic inheritance[87]. Regardless of the mechanism, the PIWI–piRNA pathway clearly functions to suppress expression of new phenotypes and to maintain developmental robustness in *Drosophila*.

The uncertain meaning of PIWI in cancer

Cancer stem cells may help to explain resistance to cancer treatment and relapse after treatment in certain forms of cancer. Thus, great interest exists in developing our understanding of the basic biology of cancer stem cells and in identifying factors that drive their stemness. A large number of studies document the ectopic expression of PIWIs in cancers. This was first reported in seminoma, a testicular germ-cell tumour in which HIWI was drastically overexpressed[32]. Related to this overexpression, HIWI mapped to a genomic region linked to seminomas and non-seminomas. Since then, PIWI expression has been shown in a variety of somatic cancers: HIWI is expressed in gastric cancer[36,88], whereas HILI is expressed in breast cancer, colon cancer, gastrointestinal stromal tumours, renal cell carcinoma and endometrial carcinoma[33]. Some early studies also suggested that PIWI expression could be used as a prognostic marker[35]. In both hepatocellular carcinoma[34] and soft-tissue sarcoma[35], tumour HIWI expression is associated with increased risk of tumour-related death. The discovery of PIWI expression in diverse forms of human somatic cancers opens up a promising area of research, especially given the well-established role of PIWIs in stem-cell maintenance and self-renewal.

Consistent with this notion, HILI enrichment was reported in a cancer cell subpopulation expressing the stemness factors OCT4 and NANOG[89]. In addition, a large variety of embryonic and developmental genes are expressed in cancers. However, this does not necessarily imply that they have a causative role in tumorigenesis. Such a conclusion will need to be based on functional studies.

At present, such functional studies are scarce. HILI overexpression in a fibroblast cell line activates STAT3 and the antiapoptotic factor BCLX, suggesting that HILI might function as an oncogene[33]. Genetic studies in *Drosophila* revealed that *piwi* mutation attenuates tumour growth in a sensitized *lethal (3) malignant brain tumour* (*l(3)mbt*)-mutant background. Other piRNA pathway genes were also upregulated in *l(3)mbt* tumours. These findings are perhaps the strongest data available at present in correlating ectopic PIWI expression with tumour growth[90].

These preliminary mechanistic data are promising, but research on PIWIs in cancer remains at an early stage and is primarily correlative. It is known that insertional mutagenesis by LINE1 elements is common in human epithelial cancers[91]. Therefore, it is possible that PIWIs are expressed in reaction to increased transposon activity in cancer, and thus act to protect the genome. Perhaps an *in vivo* overexpression model would be a good starting point to determine whether ectopic or overexpression of a PIWI protein can actually cause cancer, or is merely a consequence of tumorigenesis. Without a doubt, more large-scale, systematic research is needed before we can conclude whether or not human PIWIs have any role in cancer.

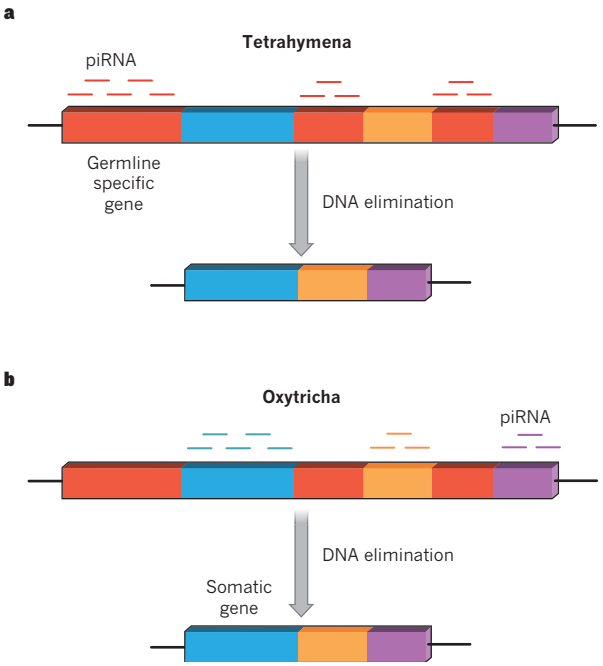


Figure 3 | Somatic genome rearrangement in ciliates. Unicellular ciliates posses two nuclei: the germline micronucleus and the somatic macronucleus. After mating, the developing macronucleus is extensively edited to remove germline-specific sequences. The new mature somatic nucleus contains only genes needed for somatic (vegetative) growth. The rest of the germline-specific genome (red boxes) undergoes DNA elimination, an extraordinary method to purge the somatic genome of repetitive sequences and transposons. This process, called somatic genome rearrangement, differs between *Tetrahymena* and *Oxytricha*. **a**, *Tetrahymena* piRNAs (red lines) are generated by the germ line, and target germline-specific sequences of the developing somatic macronucleus for elimination. **b**, *Oxytricha* piRNAs (blue, orange and purple lines) are generated by the parent somatic macronucleus, and direct the retention of somatic genes in the mature somatic macronucleus (blue, orange and purple boxes).

Outstanding questions

PIWIs occupy the interface between stem-cell and small RNA biology. They serve diverse roles in diverse tissues, from totipotent stem cells to totally differentiated cell types, from the germ line to the soma. Tantalizing questions remain, including how broadly are PIWIs and piRNAs expressed in mammalian somatic tissues? And, what is their function there, if any? The mammalian soma awaits our rigorous interrogation. At a more fundamental level, what are the molecular mechanisms by which Piwi regulates stemness? Conversely, what exactly is Piwi doing in differentiated tissues? The lower eukaryotes, in which we are making rapid progress on this topic, provide an excellent arena for these investigations.

The reason underlying ectopic expression of PIWIs in many human cancers is still a mystery. Do PIWIs have a role in cancer? If so, do PIWIs provoke dedifferentiation or lend a competitive advantage by enhancing stemness or even as a reactive mechanism to suppress transposition? Finally, are piRNAs ectopically expressed in cancers? If so, what is their function? Answers to these questions will not only shed light on the function of PIWIs and piRNAs but could also mark paths that are ripe for cancer research. ■

Received 5 August; accepted 20 November 2013.

1. Lin, H. & Spradling, A. C. A novel group of *pumilio* mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124,** 2463–2476 (1997).

2. Cox, D. N. *et al*. A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. *Genes Dev*. **12,** 3715–3727 (1998). **This paper reports the discovery of the *argonaute/piwi* gene family and is the first demonstration of the somatic function of a PIWI protein (for germline stem-cell maintenance).**
3. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442,** 199–202 (2006).
4. Aravin, A. *et al*. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442,** 203–207 (2006).
5. Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev*. **20,** 1709–1714 (2006).
6. Lau, N. C. *et al*. Characterization of the piRNA complex from rat testes. *Science* **313,** 363–367 (2006).
7. Saito, K. *et al*. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev*. **20,** 2214–2222 (2006). **This work defines a somatic piRNA pathway in the *Drosophila* ovary.**
8. Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in *Drosophila melanogaster*. *Nature* **450,** 304–308 (2007).
9. Brennecke, J. *et al*. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).
10. Aravin, A. A. *et al*. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5,** 337–350 (2003).
11. Vagin, V. V. *et al*. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
12. Deng, W. & Lin, H. *miwi*, a murine homolog of *piwi*, encodes a cytoplasmic protein essential for spermatogenesis. *Dev. Cell* **2,** 819–830 (2002).
13. Kuramochi-Miyagawa, S. *et al*. *Mili*, a mammalian member of *piwi* family gene, is essential for spermatogenesis. *Development* **131,** 839–849 (2004).
14. Carmell, M. A. *et al*. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12,** 503–514 (2007).
15. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Rev. Mol. Cell Biol*. **12,** 246–258 (2011).
16. Funayama, N., Nakatsukasa, M., Mohri, K., Masuda, Y. & Agata, K. Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges. *Evol. Dev.* **12,** 275–287 (2010).
17. Denker, E., Manuel, M., Leclere, L., Le Guyader, H. & Rabet, N. Ordered progression of nematogenesis from stem cells through differentiation stages in the tentacle bulb of *Clytia hemisphaerica* (Hydrozoa, Cnidaria). *Dev. Biol.* **315,** 99–113 (2008).
18. Seipel, K., Yanze, N. & Schmid, V. The germ line and somatic stem cell gene Cniwi in the jellyfish *Podocoryne carnea*. *Int. J. Dev. Biol*. **48,** 1–7 (2004).
19. Alié, A. *et al*. Somatic stem cells express *Piwi* and *Vasa* genes in an adult ctenophore: ancient association of "germline genes" with stemness. *Dev. Biol.* **350,** 183–197 (2011).
20. De Mulder, K. *et al*. Stem cells are differentially regulated during development, regeneration and homeostasis in flatworms. *Dev. Biol*. **334,** 198–212 (2009).
21. Reddien, P. W., Oviedo, N. J., Jennings, J. R., Jenkin, J. C. & Sanchez Alvarado, A. SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310,** 1327–1330 (2005).
22. Palakodeti, D., Smielewska, M., Lu, Y. C., Yeo, G. W. & Graveley, B. R. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* **14,** 1174–1186 (2008).
23. Rinkevich, Y. *et al*. Piwi positive cells that line the vasculature epithelium, underlie whole body regeneration in a basal chordate. *Dev. Biol.* **345,** 94–104 (2010).
24. Rinkevich, Y. *et al.* Repeated, long-term cycling of putative stem cells between niches in a basal chordate. *Dev. Cell* **24,** 76–88 (2013).
25. Cox, D. N., Chao, A. & Lin, H. *piwi* encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127,** 503–514 (2000).
26. Brower-Toland, B. *et al*. *Drosophila Piwi* associates with chromatin and interacts directly with HP1a. *Genes Dev*. **21,** 2300–2311 (2007). **This paper shows the direct interaction between Piwi and HP1a, the binding of Piwi to chromosomes in *Drosophila* somatic cells and the epigenetic effect of such interaction and binding.**
27. Perrat, P. N. *et al*. Transposition-driven genomic heterogeneity in the *Drosophila* brain. *Science* **340,** 91–95 (2013).
28. Ghildiyal, M. *et al*. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* **320,** 1077–1081 (2008).
29. Mani, S. R., Megosh, H. & Lin, H. PIWI proteins are essential for early *Drosophila* embryogenesis. *Dev. Biol*. http://dx.doi.org/10.1016/j.ydbio.2013.10.017 (31 October 2013).
30. Sharma, A. K. *et al*. Human CD34$^+$ stem cells express the *hiwi* gene, a human homologue of the *Drosophila* gene *piwi*. *Blood* **97,** 426–434 (2001).
31. Nolde, M. J., Cheng, E. C., Guo, S. & Lin, H. *Piwi* genes are dispensable for normal hematopoiesis in mice. *PLoS ONE* **8,** e71950 (2013).
32. Qiao, D., Zeeman, A. M., Deng, W., Looijenga, L. H. & Lin, H. Molecular characterization of *hiwi*, a human member of the *piwi* gene family whose overexpression is correlated to seminomas. *Oncogene* **21,** 3988–3999 (2002).
33. Lee, J. H. *et al*. Stem-cell protein Piwil2 is widely expressed in tumors and inhibits apoptosis through activation of Stat3/Bcl-XL pathway. *Hum. Mol. Genet.* **15,** 201–211 (2006).
34. Zhao, Y. M. *et al. HIWI* is associated with prognosis in patients with hepatocellular carcinoma after curative resection. *Cancer* **118,** 2708–2717 (2012).
35. Taubert, H. *et al*. Expression of the stem cell self-renewal gene *Hiwi* and risk of tumour-related death in patients with soft-tissue sarcoma. *Oncogene* **26,** 1098–1100 (2007).
36. Liu, X. *et al*. Expression of *hiwi* gene in human gastric cancer was associated with proliferation of cancer cells. *Int. J. Cancer* **118,** 1922–1929 (2006).
37. Yan, Z. *et al*. Widespread expression of piRNA-like molecules in somatic tissues. *Nucleic Acids Res*. **39,** 6596–6607 (2011).
38. Malone, C. D. *et al*. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137,** 522–535 (2009). **This work delineates distinct germline and somatic piRNA pathways in the *Drosophila* ovary.**
39. Saito, K. *et al*. A regulatory circuit for *piwi* by the large *Maf* gene traffic jam in *Drosophila*. *Nature* **461,** 1296–1299 (2009).
40. Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L. & Hannon, G. J. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature* **491,** 279–283 (2012).
41. Olivieri, D., Sykora, M. M., Sachidanandam, R., Mechtler, K. & Brennecke, J. An *in vivo* RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in *Drosophila*. *EMBO J*. **29,** 3301–3317 (2010).
42. Saito, K. *et al*. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in *Drosophila*. *Genes Dev*. **24,** 2493–2498 (2010).
43. Qi, H. *et al.* The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. *J. Biol. Chem*. **286,** 3789–3797 (2011).
44. Szakmary, A., Reedy, M., Qi, H. & Lin, H. The Yb protein defines a novel organelle and regulates male germline stem cell self-renewal in *Drosophila melanogaster*. *J. Cell Biol*. **185,** 613–627 (2009).
45. Handler, D. *et al*. A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO J.* **30,** 3977–3993 (2011).
46. Nishimasu, H. *et al*. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature* **491,** 284–287 (2012).
47. Watanabe, T. *et al*. MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. *Dev. Cell* **20,** 364–375 (2011).
48. Zheng, K. *et al*. Mouse MOV10L1 associates with Piwi proteins and is an essential component of the Piwi-interacting RNA (piRNA) pathway. *Proc. Natl Acad. Sci. USA* **107,** 11841–11846 (2010).
49. Frost, R. J. *et al*. MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs. *Proc. Natl Acad. Sci. USA* **107,** 11847–11852 (2010).
50. Xiol, J. *et al*. A role for Fkbp6 and the chaperone machinery in piRNA amplification and transposon silencing. *Mol. Cell* **47,** 970–979 (2012).
51. Crackower, M. A. *et al*. Essential role of Fkbp6 in male fertility and homologous chromosome pairing in meiosis. *Science* **300,** 1291–1295 (2003).
52. Pandey, R. R. *et al*. Tudor domain containing 12 (TDRD12) is essential for secondary PIWI interacting RNA biogenesis in mice. *Proc. Natl Acad. Sci*. **110,** 16492–16497 (2013).
53. Darricarrère, N., Liu, N., Watanabe, T. & Lin, H. Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. *Proc. Natl Acad. Sci. USA* **110,** 1297–1302 (2013).
54. Olivieri, D., Senti, K. A., Subramanian, S., Sachidanandam, R. & Brennecke, J. The cochaperone shutdown defines a group of biogenesis factors essential for all piRNA populations in *Drosophila*. *Mol. Cell* **47,** 954–969 (2012).
55. Kawaoka, S., Izumi, N., Katsuma, S. & Tomari, Y. 3′ end formation of PIWI-interacting RNAs *in vitro*. *Mol. Cell* **43,** 1015–1022 (2011).
56. Parker, J. S., Roe, S. M. & Barford, D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J*. **23,** 4727–4737 (2004).
57. Boland, A., Huntzinger, E., Schmidt, S., Izaurralde, E. & Weichenrieder, O. Crystal structure of the MID-PIWI lobe of a eukaryotic Argonaute protein. *Proc. Natl Acad. Sci. USA* **108,** 10466–10471 (2011).
58. Nakanishi, K., Weinberg, D. E., Bartel, D. P. & Patel, D. J. Structure of yeast Argonaute with guide RNA. *Nature* **486,** 368–374 (2012).
59. Saito, K., Sakaguchi, Y., Suzuki, T., Siomi, H. & Siomi, M. C. Pimet, the *Drosophila* homolog of HEN1, mediates 2′-O-methylation of Piwi- interacting RNAs at their 3′ ends. *Genes Dev.* **21,** 1603–1608 (2007).
60. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Mol. Cell* **9,** 315–327 (2002). **This is the first demonstration that Piwi is essential for transcriptional and post-transcriptional gene silencing.**
61. Grimaud, C. *et al*. RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124,** 957–971 (2006).
62. Pal-Bhadra, M. *et al*. Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303,** 669–672 (2004).
63. Sienski, G., Donertas, D. & Brennecke, J. Transcriptional silencing of transposons by Piwi and Maelstrom and its impact on chromatin state and gene expression. *Cell* **151,** 964–980 (2012).
64. Huang, X. A. *et al*. A major epigenetic programming mechanism guided by piRNAs. *Dev. Cell* **24,** 502–516 (2013). **This work demonstrates that piRNAs are both necessary and sufficient to recruit Piwi and epigenetic factors to target sites and presents a whole-genome analysis of Piwi binding.**
65. Le Thomas, A. *et al*. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev*. **27,** 390–399 (2013).

66. Rozhkov, N. V., Hammell, M. & Hannon, G. J. Multiple roles for Piwi in silencing *Drosophila* transposons. *Genes Dev.* **27,** 400–412 (2013).
67. Ge, D. T. & Zamore, P. D. Small RNA-directed silencing: the fly finds its inner fission yeast? *Curr. Biol.* **23,** R318–R320 (2013).
68. Muerdter, F. *et al*. A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in *Drosophila*. *Mol. Cell* **50,** 736–748 (2013).
69. Ohtani, H. *et al*. DmGTSF1 is necessary for Piwi-piRISC-mediated transcriptional transposon silencing in the *Drosophila* ovary. *Genes Dev.* **27,** 1656–1661 (2013).
70. Dönertas, D., Sienski, G. & Brennecke, J. *Drosophila* Gtsf1 is an essential component of the Piwi-mediated transcriptional silencing complex. *Genes Dev.* **27,** 1693–1705 (2013).
71. Akkouche, A. *et al*. Maternally deposited germline piRNAs silence the *tirant* retrotransposon in somatic cells. *EMBO Rep.* **14,** 458–464 (2013).
72. Gu, T. & Elgin, S. C. R. Maternal depletion of Piwi, a component of the RNAi system, impacts heterochromatin formation in *Drosophila*. *PLoS Genet.* **9,** e1003780 (2013).
73. Schoeberl, U. E. & Mochizuki, K. Keeping the soma free of transposons: programmed DNA elimination in ciliates. *J. Biol. Chem.* **286,** 37045–37052 (2011).
74. Mochizuki, K. & Gorovsky, M. A. A dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev.* **19,** 77–89 (2005).
75. Okamura, K. & Lai, E. C. Endogenous small interfering RNAs in animals. *Nature Rev. Mol. Cell Biol.* **9,** 673–678 (2008).
76. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *tetrahymena*. *Cell* **110,** 689–699 (2002).
This demonstrates that a PIWI protein is required for DNA elimination/ rearrangement in *Tetrahymena*.
77. Taverna, S. D., Coyne, R. S. & Allis, C. D. Methylation of histone h3 at lysine 9 targets programmed DNA elimination in *tetrahymena*. *Cell* **110,** 701–711 (2002).
78. Aronica, L. *et al*. Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in *Tetrahymena*. *Genes Dev.* **22,** 2228–2241 (2008).
79. Fang, W., Wang, X., Bracht, J. R., Nowacki, M. & Landweber, L. F. Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell* **151,** 1243–1255 (2012).
80. Rajasethupathy, P. *et al*. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149,** 693–707 (2012).
This work reports an epigenetic function for the PIWI–piRNA pathway in the *Aplysia* neural system.
81. Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150,** 563–565 (1942).
82. Queitsch, C., Sangster, T. A. & Lindquist, S. Hsp90 as a capacitor of phenotypic variation. *Nature* **417,** 618–624 (2002).
83. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396,** 336–342 (1998).
84. Gangaraju, V. K. *et al*. *Drosophila* Piwi functions in Hsp90-mediated suppression of phenotypic variation. *Nature Genet.* **43,** 153–158 (2011).
85. Specchia, V. *et al*. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature* **463,** 662–665 (2010).
86. Sollars, V. *et al.* Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nature Genet.* **33,** 70–74 (2003).
87. Ashe, A. *et al.* piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. *Cell* **150,** 88–99 (2012).
88. Grochola, L. F. *et al*. The stem cell-associated *Hiwi* gene in human adenocarcinoma of the pancreas: expression and risk of tumour-related death. *Br. J. Cancer* **99,** 1083–1088 (2008).
89. Lee, J. H. *et al*. Pathways of proliferation and antiapoptosis driven in breast cancer stem cells by stem cell protein Piwil2. *Cancer Res.* **70,** 4569–4579 (2010).
90. Janic, A., Mendizabal, L., Llamazares, S., Rossell, D. & Gonzalez, C. Ectopic expression of germline genes drives malignant brain tumor growth in *Drosophila*. *Science* **330,** 1824–1827 (2010).
91. Lee, E. *et al*. Landscape of somatic retrotransposition in human cancers. *Science* **337,** 967–971 (2012).
92. Harris, A. N. & Macdonald, P. M. Aubergine encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. *Development* **128,** 2823–2832 (2001).
93. Li, C. *et al.* Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137,** 509–521 (2009).

Acknowledgements We thank S. Christensen, C. Juliano and S. Ramesh Mani for their critical reading of the manuscript. R.J.R. and M.M.W. are supported by an NIH Medical Scientist Training Program grant (T32-GM07205). The current work in the Lin lab on PIWIs and piRNA is supported by the NIH (DP1CA174418 and R01HD42012), the G. Harold & Leila Mathers Foundation, and an Ellison Medical Foundation Senior Scholar Award.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/mjff5j. Correspondence should be addressed to H.L. (haifan.lin@yale.edu).

ARTICLE

doi:10.1038/nature12818

CNVs conferring risk of autism or schizophrenia affect cognition in controls

Hreinn Stefansson[1]*, Andreas Meyer-Lindenberg[2]*, Stacy Steinberg[1], Brynja Magnusdottir[3], Katrin Morgen[2], Sunna Arnarsdottir[1,3], Gyda Bjornsdottir[1], G. Bragi Walters[1], Gudrun A. Jonsdottir[1], Orla M. Doyle[4], Heike Tost[2], Oliver Grimm[2], Solveig Kristjansdottir[1], Heimir Snorrason[1], Solveig R. Davidsdottir[3], Larus J. Gudmundsson[1], Gudbjorn F. Jonsson[1], Berglind Stefansdottir[1], Isafold Helgadottir[3], Magnus Haraldsson[3,5], Birna Jonsdottir[6], Johan H. Thygesen[7], Adam J. Schwarz[8], Michael Didriksen[9], Tine B. Stensbøl[9], Michael Brammer[4], Shitij Kapur[4], Jonas G. Halldorsson[5], Stefan Hreidarsson[10], Evald Saemundsen[5,10], Engilbert Sigurdsson[3,5] & Kari Stefansson[1]

In a small fraction of patients with schizophrenia or autism, alleles of copy-number variants (CNVs) in their genomes are probably the strongest factors contributing to the pathogenesis of the disease. These CNVs may provide an entry point for investigations into the mechanisms of brain function and dysfunction alike. They are not fully penetrant and offer an opportunity to study their effects separate from that of manifest disease. Here we show in an Icelandic sample that a few of the CNVs clearly alter fecundity (measured as the number of children by age 45). Furthermore, we use various tests of cognitive function to demonstrate that control subjects carrying the CNVs perform at a level that is between that of schizophrenia patients and population controls. The CNVs do not all affect the same cognitive domains, hence the cognitive deficits that drive or accompany the pathogenesis vary from one CNV to another. Controls carrying the chromosome 15q11.2 deletion between breakpoints 1 and 2 (15q11.2(BP1–BP2) deletion) have a history of dyslexia and dyscalculia, even after adjusting for IQ in the analysis, and the CNV only confers modest effects on other cognitive traits. The 15q11.2(BP1–BP2) deletion affects brain structure in a pattern consistent with both that observed during first-episode psychosis in schizophrenia and that of structural correlates in dyslexia.

Little information is available on whether or how rare CNVs conferring high risk of schizophrenia and/or autism affect physiologic function of otherwise normal brains. As none of these CNVs hitherto described are fully penetrant for the diseases, and both schizophrenia and autism affect cognition, we aimed to examine the possibility that the CNVs affect cognition in control carriers, those who do not suffer either disease or intellectual disability. We based our selection of CNVs on a literature search for CNVs associated with schizophrenia and/or autism ('neuropsychiatric CNVs'); this search produced 26 CNV alleles (Supplementary Table 1)[1–3]. These CNV alleles are rare, found in 0.002% to 0.2% frequency, and cumulatively in 1.16% of our sample of 101,655 genotyped subjects, representing approximately one-third of the Icelandic population (Supplementary Tables 1 and 2).

We used the subset of genotyped subjects born before 1968, without excluding patients, to examine the association of each neuropsychiatric CNV with reproductive outcome ('fecundity'), defined simply as the number of children each subject had by age 45. After correction for multiple comparisons, three neuropsychiatric CNVs were significantly associated with fecundity (Table 1). Subjects carrying the 16p11.2 deletion or the 22q11.21 duplication show reduced fecundity, with the effect in males significantly greater than in females ($P = 0.0083$ and $P = 0.029$ for the difference in effect by sex for the 16p11.2 deletion and the 22q11.21 duplication, respectively). In contrast, individuals carrying the 16p12.1 deletion have more children than do controls (Table 1). Those with deletions at 15q11.2(BP1-BP2) show a nominally significant reduction in fecundity (Table 1). Consistent with previous reports[4], schizophrenia patients show a large decrease in fecundity, with a more pronounced reduction in males ($P = 9.5 \times 10^{-25}$ for the difference in effect by sex) (Table 1).

We recruited neuropsychiatric CNV control carriers, controls carrying other CNVs not known to be associated with schizophrenia or autism ('other CNVs'), controls without large CNVs, and schizophrenia patients. All recruited subjects (Supplementary Table 2 and Supplementary Fig. 1) were administered a battery of neuropsychological tests (see Methods), the mini international neuropsychiatric interview (MINI)[5] and the general assessment of function scale (GAF)[6].

The neuropsychiatric CNVs as a class

We found that the GAF score is 0.70 standard deviations (s.d.) lower in the group of neuropsychiatric CNV control carriers than in population controls ($P = 2.2 \times 10^{-12}$). Based on MINIs, anxiety and substance abuse prevalences in the neuropsychiatric CNV control group are similar to those of controls ($P = 0.27$ and 0.36, respectively), however, depression and suicidal ideation are more common (odds ratio = 2.86, $P = 0.0017$, and odds ratio = 2.20, $P = 0.011$, respectively). The other CNVs have GAF scores 0.18 s.d. lower than population controls ($P = 0.0098$), but do not differ significantly from controls in prevalence of phenotypes assessed by the MINI ($P = 0.22$, 0.90, 0.97 and 0.097, for depression, suicidal ideation, anxiety and substance abuse, respectively).

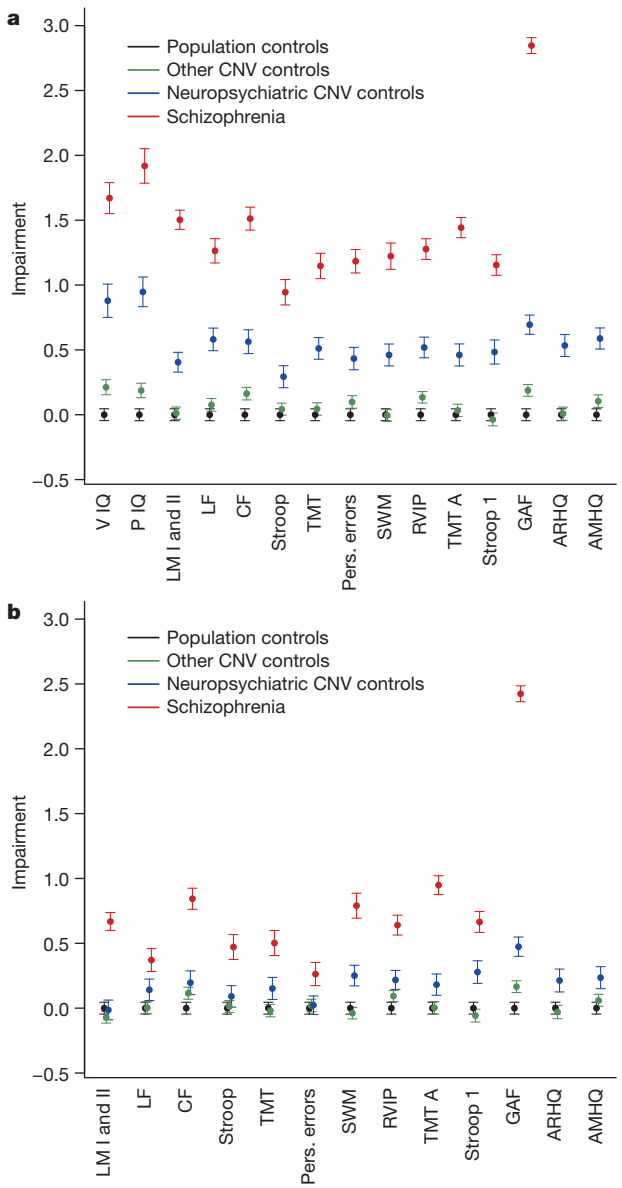
Neurocognitive deficits, or heritable neurocognitive traits, are seen in those at risk of schizophrenia and in unaffected family members[7,8]. They typically distinguish patients with schizophrenia from controls

[1]deCODE genetics/Amgen, Sturlugata 8, IS-101 Reykjavík, Iceland. [2]Central Institute of Mental Health, University of Heidelberg Medical Faculty Mannheim, 68159 Mannheim, Germany. [3]Landspitali, Department of Psychiatry, National University Hospital, IS-101 Reykjavík, Iceland. [4]Institute of Psychiatry, King's College, 16 De Crespigny Park, London SE5 8AF, UK. [5]University of Iceland, Faculty of Medicine, University of Iceland, IS-101 Reykjavík, Iceland. [6]Röntgen Domus, Egilsgötu 3, IS-101 Reykjavík, Iceland. [7]Mental Health Centre Sct. Hans, Copenhagen University Hospital, Research Institute of Biological Psychiatry, Boserupvej 2, DK-4000 Roskilde, Denmark. [8]Tailored Therapeutics, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center DC 1940, Indianapolis, Indiana 46285, USA. [9]H. Lundbeck A/S, Ottiliavej 9, DK-2500 Valby, Denmark. [10]The State Diagnostic and Counselling Centre, Digranesvegur 5, IS-200 Kópavogur, Iceland.
*These authors contributed equally to this work.

Table 1 | Fecundity of neuropsychiatric CNV carriers and schizophrenia patients based on individuals born before 1968

CNV	Carriers (male/female)	Non carriers (male/female)	Effect (male/female)	*P* value
16p11.2 del	14/13	33910/42223	0.14/0.51	1.6×10^{-12}
22q11.21 dup	25/33	33899/42203	0.60/0.90	0.00093
16p12.1 del	36	76124	1.32	0.0011
15q11.2 del	73/99	33851/42137	0.81/1.03	0.015
1q21.1 del	21	76139	0.76	0.062
15q11.2–13.1 dup	9	76151	0.69	0.11
16p13.1 dup	90	76070	0.91	0.14
16p11.2 distal del	11	76149	0.76	0.2
16p13.1 del	28	76132	0.86	0.22
17q12 dup	28	76132	1.11	0.3
2p16.3 (NRXN1) del	10	76150	0.83	0.34
16p11.2 dup	26	76134	0.89	0.35
22q11.21 del	7	76153	0.79	0.37
17p12 del	24	76136	1.08	0.48
13q31.3 (GPC6) dup	76	76084	0.96	0.55
10q11.22–23 dup	13	76147	1.08	0.62
10q11.22–23 del	12	76148	0.96	0.81
1q21.1 dup	33	76127	1.02	0.88
15q13.1 dup	17	76143	1.02	0.9
2p25.3 (MYT1L) dup	106	76054	0.99	0.91
15q13.3 all del	18	76142	1.01	0.95
Schizophrenia	306/197	33618/42039	0.21/0.54	9.5×10^{-206}

The effect is the factor by which fecundity is altered in CNV carriers or schizophrenia patients. Different effects on males and females are shown when there is a significant ($P < 0.05$) interaction between sex and either CNV or patient status. Counting both models fitted (with and without sex interaction), 42 tests were performed, thus the significance threshold for CNVs is $P = 0.0012$.



with an effect size of approximately one s.d. (ref. 9). The cognitive deficits seen in first-degree relatives of schizophrenia patients indicate that they are partly independent of the clinical state[10] and are likely to persist in psychotic patients, at least in a milder form, even if a complete remission would be achieved. All subjects were administered tests for cognitive profiling that measure functions previously shown to be impaired in schizophrenia patients, including attention, spatial working memory, logical memory, executive function, cognitive flexibility, language and processing speed (see Methods). On all tests, the schizophrenia patients performed worse than population controls (Fig. 1a and Supplementary Table 3a). The neuropsychiatric CNV control carriers performed at a level between that of schizophrenia patients and population controls, whereas the controls carrying other CNVs performed in line with population controls (Fig. 1a and Supplementary Table 3a). For all the tests, the association is much weaker when IQ is taken into account (Fig. 1b and Supplementary Table 3b). This is not surprising as the cognitive tests measure attributes that contribute to IQ.

Scores on the adult reading history questionnaire (ARHQ)[11,12] and adult mathematical history questionnaire (AMHQ) questionnaires (see Methods), designed to detect a history of reading and mathematics learning difficulties indicative of dyslexia and dyscalculia, separate the control neuropsychiatric CNV carriers from population controls with an effect of 0.50 s.d. ($P = 3.1 \times 10^{-6}$) and 0.55 s.d. ($P = 2.5 \times 10^{-7}$), respectively (Fig. 1a). ARHQ and AMHQ scores for the carriers of other CNVs are not significantly different from those of population controls ($P = 0.98$ and 0.17, respectively).

Figure 1 | Association of CNV groups with cognitive traits, GAF, ARHQ and AMHQ scores. **a**, Average standardized scores for schizophrenia patients ($n = 161$), control carriers of neuropsychiatric CNVs ($n = 167$), control carriers of other CNVs ($n = 465$) and population controls ($n = 475$). **b**, Average standardized scores after adjustment for IQ. AMHQ, adult mathematical history questionnaire; ARHQ, adult reading history questionnaire; CF, category fluency; GAF, global assessment of functioning; LF, letter fluency; LM I and II, logical memory I and II; Pers. errors, perseverative errors; P IQ, performance IQ; RVIP, rapid visual information processing; Stroop, difference in time to complete trial 3 and time to complete trial 2; Stroop 1, Stroop trial 1; SWM, spatial working memory (between-search errors for 6 boxes); TMT, TMT trail B − TMT trail A; TMT A, TMT trail A; V IQ, verbal IQ; (see Methods for further information on tests). Error bars represent s.e.m. Impairment is in s.d. units, ARHQ and AMHQ scores for the patient group are not available.

Individual neuropsychiatric CNVs

To determine whether the CNVs differ in their effects on cognition, we examined the association of individual CNVs with neurocognitive traits, GAF score and history of learning difficulties. Few control carriers could be recruited for some of the neuropsychiatric CNVs but between 5 and 47 control carriers could be evaluated for each of 11 CNVs (Supplementary Table 2). Six of the CNVs are associated with verbal and/or performance IQ with large effects (0.73–3.51 s.d. units) in the carrier controls. These are the 16p11.2 deletion and the reciprocal duplication, 17p12 deletion, 17q12 duplication, 16p12.1 deletion and 16p13.1 duplication (Table 2). The effect is also large for the 2p16.3 deletion carriers for performance IQ, although the P value is >0.005 (Supplementary Table 4a). Significant associations were also found between individual neuropsychiatric CNVs and GAF, spatial working memory (SWM), AMHQ, category fluency, letter fluency, perseverative errors and Stroop trial 1 (Table 2).

The alleles of the 16p11.2 CNV confer mirrored effects on anthropometric traits[13]. The deletion, conferring high risk of autism[14], shows the greatest impairments in the cognitive domains tested in the control carriers. The reciprocal duplication, conferring risk of schizophrenia[15] and autism[14], confers somewhat different abnormalities on the control carriers (Supplementary Table 5). Although the deletion is strongly associated with impaired verbal IQ and deficits in verbal letter and category fluency tests, in keeping with what is seen in autism, the duplication more selectively impairs the spatial working memory and executive functions that seem to be more important in the pathophysiology of schizophrenia[16].

Four neuropsychiatric CNVs, duplications at 13q31.3, 22q11.21 and 1q21.1 and a deletion at 15q11.2(BP1-BP2), show more modest (around 0.5 s.d. or less) or no effects on verbal and performance IQ (Supplementary Table 4a). Twenty-one control subjects carrying the 22q11.21 duplication were evaluated and trends were seen for impairments in all neurocognitive traits and the most significant impairment was observed in category fluency (0.97 s.d., $P = 1.4 \times 10^{-4}$) (Supplementary Table 4a). Ten control subjects carrying the 1q21.1 duplication were evaluated and not even a nominally significant effect was detected on neurocognitive traits, GAF or history of learning difficulties (Supplementary Table 4a). Forty-seven control subjects carrying the 15q11.2(BP1-BP2) deletion were evaluated, and significant associations were observed with a lower GAF score (0.66 s.d., $P = 9.9 \times 10^{-5}$), history of learning difficulties as evaluated by the ARHQ (0.70 s.d., $P = 1.9 \times 10^{-4}$) and the AMHQ (0.78 s.d., $P = 2.3 \times 10^{-5}$) (Supplementary Table 4a). Association with a lower GAF score indicates impaired functioning, possibly due to some psychological disturbance, although the number of carriers studied was too small to allow detection of association with the individual phenotypes as derived from the MINI.

When conditioned on IQ the associations with specific cognitive traits, GAF and history of learning difficulties become less significant for the 11 CNVs (Table 2 and Supplementary Table 4a). However, the association of AMHQ score with the 15q11.2(BP1-BP2) deletion remains the most significant (0.70 s.d., $P = 2.3 \times 10^{-4}$). In Fig. 2a and Supplementary Table 6 the neuropsychiatric CNV carriers are divided into those carrying the 15q11.2(BP1-BP2) and those carrying other neuropsychiatric CNVs. A clear difference in the effect of conditioning the ARHQ and AMHQ scores on IQ is observed in these two groups: in the 15q11.2(BP1-BP2) deletion group, the association with ARHQ and AMHQ scores is only slightly weakened when conditioned on IQ, whereas in the group of remaining neuropsychiatric CNV carriers (without the 15q11.2(BP1-BP2) deletion carriers) there is no longer any significant association with the history of learning difficulties after conditioning on IQ (Fig. 2b and Supplementary Table 6).

The 15q11.2(BP1-BP2) deletion has previously been shown to confer modest risk of schizophrenia[1], behavioural disturbances[17], developmental and language delay[18], and epilepsy[19]. We show that the 15q11.2(BP1-BP2) deletion has only modest impact on results of the neuropsychological tests but is still strongly associated with a history of difficulties in learning mathematics and reading (Fig. 2). IQ is only marginally lower in the controls carrying the 15q11.2(BP1-BP2) deletion than in the population controls. Using a score of greater than 0.43 on the ARHQ[11] as a surrogate for dyslexia[20], the 15q11.2(BP1-BP2) deletion is associated with dyslexia with an odds ratio of 3.18 ($P = 0.0017$). Of three previously described ARHQ subscales, based on factor analysis[11], the ARHQ

Table 2 | Controls carrying different neuropsychiatric CNVs perform worse than population controls on cognitive tests, GAF and history of learning difficulties

CNV	Cognitive trait	Effect	*P*-value	Effect (adjusted for IQ)	*P*-value (adjusted for IQ)
16p11.2 del	V IQ	3.51	5.90×10^{-16}	NA	NA
17p12.del	V IQ	2.99	2.30×10^{-9}	NA	NA
16p11.2 del	LF	2.00	2.00×10^{-7}	0.61	0.14
16p11.2 del	P IQ	2.01	1.30×10^{-6}	NA	NA
16p11.2 del	CF	1.83	2.00×10^{-6}	0.58	0.16
16p11.2 del	Stroop 1	1.8	2.80×10^{-6}	1.14	0.006
16p12.1 del	V IQ	2.05	8.30×10^{-6}	NA	NA
16p13.1 dup	P IQ	1.09	9.30×10^{-6}	NA	NA
16p11.2 del	Pers. errors	1.77	2.00×10^{-5}	0.48	0.25
15q11.2 del	AMHQ	0.78	2.30×10^{-5}	0.70	0.00023
16p11.2 dup	SWM	1.72	3.20×10^{-5}	1.51	0.00025
17q12 dup	GAF	1.63	5.10×10^{-5}	1.43	0.00037
16p11.2 del	GAF	1.55	5.80×10^{-5}	0.58	0.10
17q12 dup	V IQ	1.57	8.10×10^{-5}	NA	NA
15q11.2 del	GAF	0.66	9.90×10^{-5}	0.57	0.0012
16p11.2 del	SWM	1.49	0.00011	0.45	0.27
22q11.21 dup	CF	0.97	0.00014	0.81	0.0016
15q11.2 del	ARHQ	0.7	0.00019	0.60	0.0018
17p12 del	GAF	1.67	0.00031	1.11	0.021
16p11.2 del	TMT A	1.37	0.0004	0.4	0.33
17p12 del	Stroop	1.61	0.00043	1.13	0.018
16p11.2 dup	P IQ	1.29	0.00062	NA	NA
16p11.2 dup	TMT	1.27	0.00073	0.91	0.016
17p12 del	CF	1.48	0.0012	0.62	0.2
16p12.1 del	P IQ	1.41	0.0021	NA	NA
16p13.1 dup	V IQ	0.73	0.0022	NA	NA
17q12 dup	LF	1.2	0.0026	0.78	0.051
16p12.1 del	LM I and LM II	1.25	0.0027	0.77	0.092
16p13.1 dup	SWM	0.66	0.0049	0.54	0.026

Abbreviations for the different tests are given in the supplementary text. The significance threshold for the 11 CNVs each compared for 15 tests and 13 IQ-adjusted tests is $P = 0.00016$. AMHQ, adult mathematical history questionnaire; ARHQ, adult reading history questionnaire; CF, category fluency; GAF, general assessment of function scale; LF, letter fluency; LM I and LM II, logical memory I and II; NA, not applicable; Pers errors, perseverative errors; P IQ, performance IQ; Stroop, difference in time to complete trial 3 and time to complete trial 2; Stroop 1, Stroop trial 1; SWM, spatial working memory; TMT A, TMT trail A; V IQ, verbal IQ.

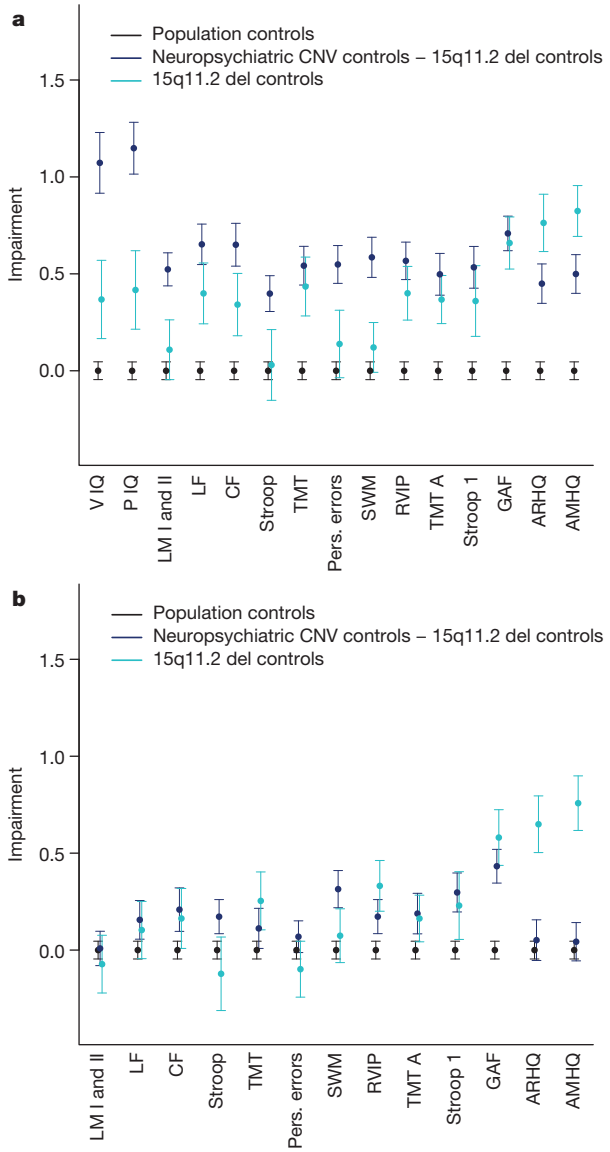


Figure 2 | Association of CNVs with cognitive traits, GAF, ARHQ and AMHQ scores. **a**, Average standardized scores for controls carrying neuropsychiatric CNVs excluding the 15q11.2(BP1-BP2) deletion carriers (blue, $n = 120$), controls carrying the15q11.2(BP1-BP2) deletion (cyan, $n = 47$) and population controls (black, $n = 475$). **b**, Average standardized scores conditioned on IQ. Error bars represent s.e.m.

dyslexia-symptoms subscale shows the strongest association with the 15q11.2(BP1-BP2) deletion (effect = 0.71 s.d., $P = 1.4 \times 10^{-4}$, Supplementary Table 7). Based on a score of greater than 12 on the AMHQ (see Methods), the 15q11.2(BP1-BP2) deletion shows association with dyscalculia (odds ratio = 3.91, $P = 0.00011$). In 136 controls carrying the reciprocal 15q11.2(BP1-BP2) duplication the results are comparable to those of population controls on the neurocognitive tests (Supplementary Table 4a), and their GAF (0.01 s.d., $P = 0.95$), ARHQ (−0.22 s.d., $P = 0.057$) and AMHQ (0.07 s.d., $P = 0.52$) scores are also in keeping with those of population controls.

Structural MRI phenotypes

In a search for neural intermediate phenotypes, we performed structural magnetic resonance imaging (MRI) on 15 control carriers of the 15q11.2(BP1-BP2) deletion, 55 carriers of the reciprocal duplication, and 201 population controls. Given the association with schizophrenia, we focused our attention on regions defined by a recent meta-analysis of first-episode psychosis[21] (Fig. 3a). The 15q11.2(BP1-BP2) deletion carriers have a reduced volume of grey matter in the perigenual anterior cingulate cortex (pACC) (Fig. 3b) and the left insula (Fig. 3c). Furthermore, deletion carriers showed reductions in white matter of the temporal lobe bilaterally and an increase in the volume of the corpus callosum (Fig. 3d). Importantly, for both grey and white matter, 15q11.2(BP1-BP2) duplication carriers always show reciprocal changes in exactly the same regions altered in deletion carriers, providing the first demonstration of allele-dose-dependent effects of CNVs on the structure of the human brain.

The pACC is a key region for regulation of limbic activity[22] previously shown to be abnormal in schizophrenia[23]. Furthermore, the frontoinsular cortex is highly connected to the pACC, with which it forms the cortical aspects of the salience network[24], a circuitry linked to schizophrenia risk[25]. Although reduction in the volume of the temporal lobe white matter is a well-established feature of schizophrenia and is present early in the illness[26], the finding of increased callosal volume was unexpected, as patients with schizophrenia have reduced volume in this region[27]. Notably, carriers of the schizophrenia-associated 22q11.21 deletion also show increased volume of the corpus callosum[28].

The abnormalities found in the structural MRI studies also show overlap with published work on structural correlates of dyslexia and dyscalculia. In dyslexia, grey matter abnormalities in the supramarginal gyrus were prominent in a recent meta-analysis[29]. Grey matter reductions in a very similar location in pACC have also been seen in developmental dyscalculia[30]. In both cognitive developmental disorders, other regions are abnormal that are not implicated in the present study (such as left perisylvian areas in dyslexia and the intraparietal sulcus in dyscalculia), suggesting that the neuropsychological impairment seen in 15q11.2(BP1-BP2) deletion carriers may be related to specific key nodes in the networks associated with these cognitive dysfunctions.

It is of interest that the controls carrying the 15q11.2(BP1-BP2) duplication ($n = 136$) perform to a similar level as population controls on all tests of cognitive function used in this study (Supplementary Table 4a). Thus, although mirror effects on brain volume phenotypes are observed for the 15q11.2(BP1-BP2) CNV, we do not observe clear mirror effects on the ARHQ and AMHQ scores.

Conclusion

There were two main aims to this study. The first was to determine whether carriers of CNVs that predispose to schizophrenia and/or autism, who have not been diagnosed with a psychotic disorder or autism, have cognitive abnormalities that are akin to those encountered in schizophrenia. If this were the case these neuropsychiatric CNVs could be used as instruments in the study of cognitive abnormalities that characterize the disease. The results show that carriers of these CNVs show cognitive abilities in between those of normal controls and patients with schizophrenia. This raises the possibility that the difference between the patients and the control carriers may not be due to a lack of penetrance but instead to variation in expressivity of the CNVs. It also shows that the cognitive abnormalities are not necessarily consequences of the disease, and that the risk of the disease may, at least in part, be mediated through the cognitive abnormalities. These CNVs could be used to identify individuals in whom schizophrenia-like cognitive abnormalities could be studied without the confounding effects of psychosis or medications.

The second aim was to better define the cognitive abnormalities in population controls carrying CNVs associating with schizophrenia and autism; by evaluating controls carrying the neuropsychiatric CNVs we sought to learn more precisely which cognitive abnormalities put carriers at risk of developing schizophrenia. We tested the carriers primarily for those aspects of cognition that have been shown to be abnormal in schizophrenia (Fig. 1), and in all of these aspects the control carriers were found to perform somewhere between the schizophrenia patients and the population controls. When controlled for IQ, the number of tests that separate the control neuropsychiatric CNV carriers from population controls decreases substantially, which is not surprising as these tests assess functions that are components of the IQ. Of the

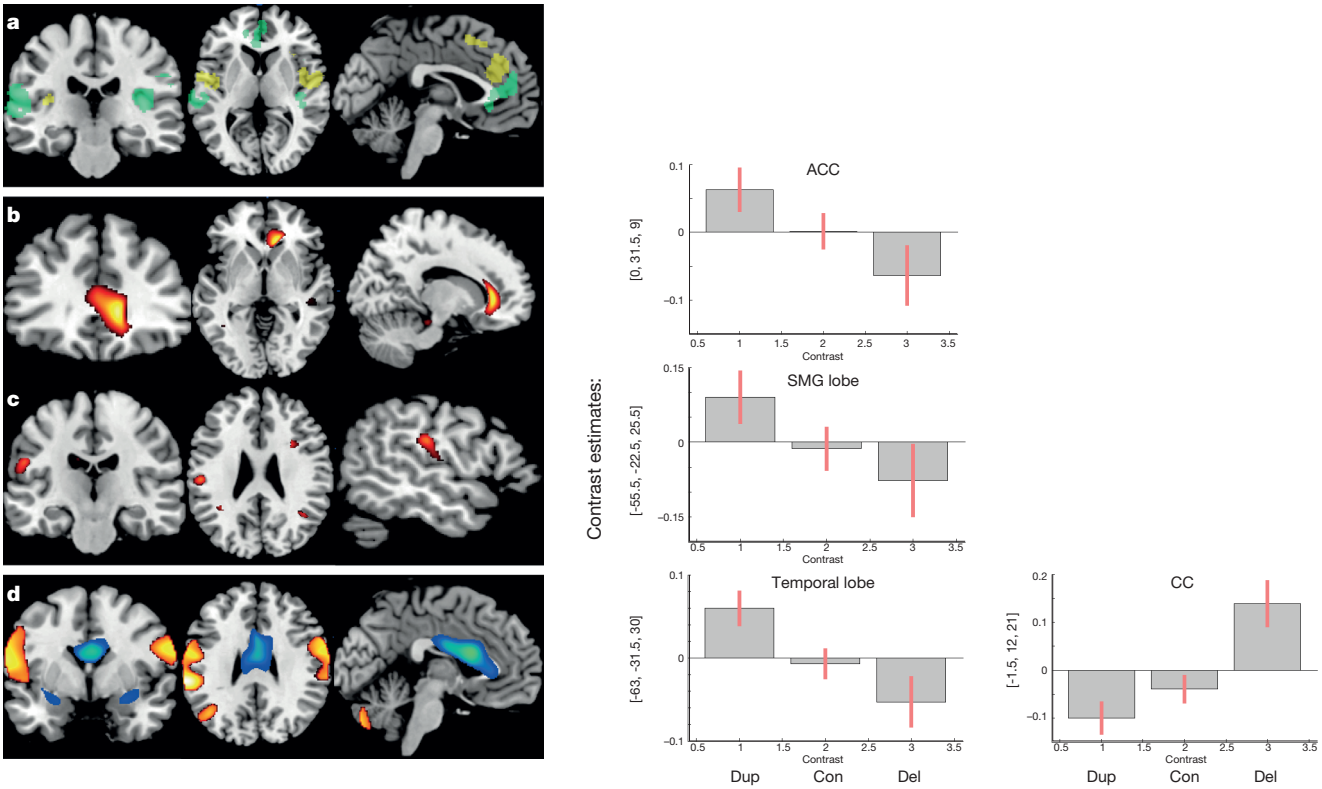


Figure 3 | Dose-dependent alterations in brain structure in 15q11.2 (BP1-BP2) CNV carriers. **a**, Regions of interest defined by a meta-analysis of first-episode psychosis in pACC and insula[21]. **b**, Grey matter alterations in perigenual anterior cingulate cortex. **c**, Grey matter alterations in insula and supramarginal gyrus (SMG). **d**, White matter alterations in left temporal lobe and corpus callosum. All data are displayed on a coronal, horizontal and sagittal section (left to right) of a structural template MRI. Maps are displayed at a threshold of $P = 0.001$ uncorrected for presentation purposes, but findings are significant at $P < 0.05$, family-wise error corrected for multiple comparisons. Right panels in **b**–**d** associate with the images to the left and show plots of contrast estimates extracted from the peak voxel in standard space (template for defining the standard space; Montreal Neuroimaging Institute). Bars represent means across subjects. Con, controls; Del, deletion; Dup, duplication. Error bars represent s.e.m.

11 neuropsychiatric CNV alleles tested independently for association in 5 or more controls, 8 associated with a cognitive trait.

Dosage-sensitive genes at CNV loci can give rise to mirrored phenotypes for anthropometric loci including body mass index[13] and head circumference[15]. In this study we provide the first evidence that dose-dependent effects of CNVs also affect human brain structure directly. Two brain regions, with clear evidence of both structural and functional alterations early in the course of schizophrenia, show dosage effects in controls carrying the 15q11.2(BP1-BP2) deletion and its reciprocal duplication.

In this paper we demonstrate how cognitive abnormalities and changes in the structure of the brain observed in schizophrenia are also found in control carriers of CNVs that confer high risk of the disease. One of the missing pieces in our understanding of the pathogenesis of schizophrenia has been the nature of the physiologic function that is first perturbed in the disease or the perturbation of which leads to the disease. We show that carriers who have not been diagnosed with autism, intellectual disability, or schizophrenia show intermediate phenotypes in brain structure that are in good agreement with the observations in first-episode psychosis. We suggest that the work presented here lends support to the idea that the cognitive abnormalities are fundamental defects in schizophrenia as they are manifest in carriers of CNVs conferring risk of the disease who do not suffer from the disease. Furthermore, in addition to the information they may provide on disease, these CNVs provide us with an opportunity to search systematically for the biochemical foundations of the cognitive differences between the carrier and non-carrier controls.

METHODS SUMMARY

Control subjects carrying CNVs or not carrying CNVs were recruited from a large genotyped sample. Subjects aged 18 to 65 years were recruited for cognitive phenotyping. The psychologists and psychiatrists evaluating all subjects were blind to genotype. To examine fecundity, a nation-wide genealogy database was used to calculate fecundity of patients and controls carrying neuropsychiatric CNVs. The MINI[5] was used to screen the controls for psychiatric disorders, and participants' overall level of functioning and their ability to carry out activities of daily living were rated using the GAF scale[31]. Memory was assessed using the logical memory subtest from Wechsler Memory Scale III (WMS-III)[32]. Verbal fluency was assessed using the controlled oral word association test (COWAT)[33] and the category naming test[34]. The Stroop test was administered as an indicator of the ability to suppress an habitual response[35]. The trail-making test (TMT) was administered as a measure of psychomotor speed and mental flexibility[36]. As a further measure of mental flexibility, including the ability to alter cognitive sets, the Wisconsin card sorting test (WCST)[37] was administered and the ratio of perseverative errors to the number of trials administered used in our analysis. Spatial working memory and sustained attention were evaluated by the computerized CANTAB battery, using the SWM[38] and rapid visual information processing (RVIP) subtests[39], respectively. Intelligence was evaluated using the Wechsler Abbreviated Scale of Intelligence (WASI-I)[40]. For neuroimaging, MRI examinations were conducted on a 1.5 T whole body Philips Achieva scanner. High-resolution T1-weighted images were processed according to the unified segmentation model with SPM8 and Matlab 8b software. Copy-number effects were examined on a voxel-by-voxel basis with a multiple regression model using SPM8.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 June; accepted 29 October 2013.
Published online 18 December 2013; corrected online 15 January 2014 (see full-text HTML version for details).

1. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).

2. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* (2013).
3. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148,** 1223–1241 (2012).
4. Haukka, J., Suvisaari, J. & Lonnqvist, J. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am. J. Psychiatry* **160,** 460–463 (2003).
5. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **59** (Suppl. 20), 22–33, quiz 34–57 (1998).
6. Meins, W., Jacobsen, G. & Stratmann, C. Social adjustment of psychiatric patients: evaluation of a modified version of the GAF (Global Assessment of Functioning) Scale. [in German with English abstract] *Psychiatr. Prax.* **22,** 206–208 (1995).
7. Sitskoorn, M. M., Aleman, A., Ebisch, S. J., Appels, M. C. & Kahn, R. S. Cognitive deficits in relatives of patients with schizophrenia: a meta-analysis. *Schizophr. Res.* **71,** 285–295 (2004).
8. Snitz, B. E., Macdonald, A. W., III & Carter, C. S. Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. *Schizophr. Bull.* **32,** 179–194 (2006).
9. Mesholam-Gately, R. I., Giuliano, A. J., Goff, K. P., Faraone, S. V. & Seidman, L. J. Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology* **23,** 315–336 (2009).
10. Rund, B. R. A review of longitudinal studies of cognitive functions in schizophrenia patients. *Schizophr. Bull.* **24,** 425–435 (1998).
11. Bjornsdottir, G. *et al.* The adult reading history questionnaire (ARHQ) in Icelandic: psychometric properties and factor structure. *J. Learn. Disabil.* (2013).
12. Lefly, D. L. & Pennington, B. F. Reliability and validity of the adult reading history questionnaire. *J. Learn. Disabil.* **33,** 286–296 (2000).
13. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478,** 97–102 (2011).
14. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
15. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet.* **41,** 1223–1227 (2009).
16. Park, S. & Holzman, P. S. Schizophrenics show spatial working memory deficits. *Arch. Gen. Psychiatry* **49,** 975–982 (1992).
17. Doornbos, M. *et al.* Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances. *Eur. J. Med. Genet.* **52,** 108–115 (2009).
18. Burnside, R. D. *et al.* Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. *Hum. Genet.* **130,** 517–528 (2011).
19. de Kovel, C. G. *et al.* Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* **133,** 23–32 (2010).
20. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems* 10th edn (WHO, 2008).
21. Radua, J. *et al.* Multimodal meta-analysis of structural and functional brain changes in first episode psychosis and the effects of antipsychotic medication. *Neurosci. Biobehav. Rev.* **36,** 2325–2333 (2012).
22. Diorio, D., Viau, V. & Meaney, M. J. The role of the medial prefrontal cortex (cingulate gyrus) in the regulation of hypothalamic-pituitary-adrenal responses to stress. *J. Neurosci.* **13,** 3839–3847 (1993).
23. Lederbogen, F. *et al.* City living and urban upbringing affect neural social stress processing in humans. *Nature* **474,** 498–501 (2011).
24. Seeley, W. W. *et al.* Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* **27,** 2349–2356 (2007).
25. Palaniyappan, L. & Liddle, P. F. Does the salience network play a cardinal role in psychosis? An emerging hypothesis of insular dysfunction. *J. Psychiatry Neurosci.* **37,** 17–27 (2012).
26. Bora, E. *et al.* Neuroanatomical abnormalities in schizophrenia: a multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophr. Res.* **127,** 46–57 (2011).
27. Arnone, D., McIntosh, A. M., Tan, G. M. & Ebmeier, K. P. Meta-analysis of magnetic resonance imaging studies of the corpus callosum in schizophrenia. *Schizophr. Res.* **101,** 124–132 (2008).
28. Tan, G. M., Arnone, D., McIntosh, A. M. & Ebmeier, K. P. Meta-analysis of magnetic resonance imaging studies in chromosome 22q11.2 deletion syndrome (velocardiofacial syndrome). *Schizophr. Res.* **115,** 173–181 (2009).
29. Linkersdörfer, J., Lonnemann, J., Lindberg, S., Hasselhorn, M. & Fiebach, C. J. Grey matter alterations co-localize with functional abnormalities in developmental dyslexia: an ALE meta-analysis. *PLoS ONE* **7,** e43122 (2012).
30. Rotzer, S. *et al.* Optimized voxel-based morphometry in children with developmental dyscalculia. *Neuroimage* **39,** 417–422 (2008).
31. Hall, R. C. Global assessment of functioning. A modified scale. *Psychosomatics* **36,** 267–275 (1995).
32. Wechsler, D. *Wechsler Memory Scale* 3rd edn (Harcourt Brace and Company, 1997).
33. Benton, A. H. K. *Multilingual Aphasia Examination* (AJA Associates, 1989).
34. Morris, J. C. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **39,** 1159–1165 (1989).
35. Stroop, J. R. Studies of interference in seral verbal reactions. *J. Exp. Psychol.* **18,** 643–662 (1935).
36. Reitan, R. M. Validity of the trail making test as an indicator of organic brain damage. *Percept. Mot. Skills* **8,** 271–276 (1958).
37. Berg, E. A. A simple objective test for measuring flexibility in thinking. *J. Gen. Psychol.* **39,** 15–22 (1948).
38. Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E. & Robbins, T. W. Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* **28,** 1021–1034 (1990).
39. Sahakian, B., Jones, G., Levy, R., Gray, J. & Warburton, D. The effects of nicotine on attention, information processing, and short-term memory in patients with dementia of the Alzheimer type. *Brit. J. Psychiatry* **154,** 797–800 (1989).
40. Wechsler, D. *Wechsler Abbreviated Scale of Intelligence* (Harcourt Brace and Company, 1999).
41. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
42. Hall, R. C. & Parks, J. The modified global assessment of functioning scale: addendum. *Psychosomatics* **36,** 416–417 (1995).
43. Golden, C. J. *The Stroop color word test* (Stoelting Company, 1978).
44. Heaton, R., Chelune, G., Talley, J., Kay, G. & Curtis, G. *Wisconsin Card Sorting Test manual* (Psychological Assessment Resources, 1993).
45. Feigenbaum, J. D., Polkey, C. E. & Morris, R. G. Deficits in spatial working memory after unilateral temporal lobectomy in man. *Neuropsychologia* **34,** 163–176 (1996).
46. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26,** 839–851 (2005).
47. Manjón, J. V., Coupe, P., Marti-Bonmati, L., Collins, D. L. & Robles, M. Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* **31,** 192–203 (2010).
48. Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38,** 95–113 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors are grateful to the participants and we thank the research nurses and staff at the Krókháls recruitment centre and roentgentechnicians at Röntgen Domus. We also thank the staff at deCODE genetics core facilities and all our colleagues for their important contribution to this work. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115008 of which resources are composed of EFPIA in-kind contribution and financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EU funded FP7-People-2011-IAPP grant PsychDPC (GA 286213).

Author Contributions H.St., A.M-L., S.S., B.M., S.A., G.B., G.B.W., M.D., T.B.S., M.B., S. Ka., J.H., S.H., E.Sa., E.Si. and K.S. were involved in study design. B.M., S.A., G.A.J., S. Kr., H.Sn., S.R.D., B.S., I.H., M.H., B.J., J.G.H., S.H., E.Sa. and E.Si. were involved with cohort ascertainment, phenotypic characterization and recruitment. H.St., K.M., G.B., G.B.W., O.M.D., H.T., O.G., G.F.J., J.H.T and L.J.G. were involved with informatics and data management. H.St., A.M-L., S.S., K.M., G.B., G.B.W., O.M.D., H.T., O.G., M.B. and A.J.S. carried out statistical analysis. H.St., A.M-L., S.S., B.M., K.M., S.A., G.B., G.B.W., G.A.J., O.M.D., H.T., O.G., S. Kr., H.Sn., S.R.D., L.J.G., G.F.J., B.S., I.H., M.H., B.J., J.H.T., M.D., T.B.S., M.B., S. Ka., J.G.H., S.H., E.Sa., E.Si. and K.S. wrote the manuscript.

Author Information The authors declare competing financial interests: details are available in the online version of the paper. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.S. (kstefans@decode.is) or A.M-L. (A.Meyer-Lindenberg@zi-mannheim.de).

METHODS

Control subjects carrying or not carrying CNVs were recruited from a large genotyped sample. Subjects aged between 18 and 65 years were recruited for cognitive phenotyping. The psychologists evaluating all subjects were blind to genotype. A nation-wide genealogy database was used to calculate the fecundity of patients and controls carrying neuropsychiatric CNVs. The mini international neuropsychiatric interview (MINI)[5] was used to screen the controls for psychiatric disorders, and participants' overall level of functioning and their ability to carry out activities of daily living were rated using the general assessment of function (GAF) Scale[31]. Memory was assessed using the logical memory subtest from Wechsler Memory Scale III (WMS-III)[32]. Verbal fluency was assessed using the controlled oral word association test (COWAT)[33] and the category naming test[34]. The Stroop test was administered as an indicator of the ability to suppress an habitual response[35]. The trail-making test (TMT) was administered as a measure of psychomotor speed and mental flexibility[36]. As a further measure of mental flexibility, including the ability to alter cognitive sets, the Wisconsin card-sorting test (WCST)[37] was administered and the ratio of perseverative errors to the number of trials administered used in our analysis. Spatial working memory and sustained attention were evaluated by the computerized CANTAB battery, using the spatial working memory (SWM)[38] and rapid visual information processing (RVP) subtests[39], respectively. Intelligence was evaluated using the Wechsler Abbreviated Scale of Intelligence (WASI-I)[40]. For neuroimaging, MRI examinations were conducted on a 1.5 T whole body Philips Achieva scanner. High-resolution T1-weighted images were processed according to the unified segmentation model with SPM8 and Matlab 8b software. Copy number effects were examined on a voxel-by-voxel basis with a multiple-regression model using SPM8.

Sample. Twenty-six CNVs conferring risk of psychiatric disorders ('neuropsychiatric CNVs'), of which most are recurrent, were identified through literature search (Supplementary Table 1). Control subjects carrying neuropsychiatric CNVs were identified from a large genotyped sample ($n = 101{,}655$). The sample had been genotyped by Illumina HumanHap (300, 370, 610, 1M, 2.5M) and Illumina Omni (670, 1M, 2.5M, Express) SNP arrays. BeadStudio (Illumina; version 2.0) was used to call genotypes, normalize signal intensity data and establish the log R ratio and B allele frequency at every SNP. Samples passing quality control were examined using PennCNV[41]. All putative neuropsychiatric CNVs and other CNVs not known to be associated with schizophrenia or autism ('other CNVs') were visually inspected using DosageMiner software (developed by deCODE genetics). The neuropsychiatric CNVs, with one exception, span more than 15 SNPs on the Illumina arrays (Supplementary Table 1).

Both the neuropsychiatric CNVs and the other CNVs are large, on average around 1.5 and 1.0 Mb, respectively. All 26 neuropsychiatric CNVs delete or duplicate exons of genes, whereas 89 of the 94 other CNVs delete or duplicate exons of genes (Supplementary Table 8).

In the sample of 101,655 genotyped subjects we identified 1,178 subjects carrying one or more of the neuropsychiatric CNVs (1.16% of the sample). Carriers aged between 18 and 65 were recruited for further phenotyping and were excluded from control groups if any of the following applied: if they were diagnosed with schizophrenia, schizoaffective or bipolar disorder; if they were diagnosed with autism, intellectual disability or developmental delay at the State Diagnostic and Counseling Centre in Iceland serving children and adolescents with a disability; if they met psychoses criteria on the MINI interview; if they were diagnosed with schizophrenia, schizoaffective, bipolar disorder, autism, intellectual disability or developmental delay according to self reports (or reports from parents); if they were using antipsychotic drugs.

Phenotyped control subjects passing the exclusion criteria were: 167 controls carrying neuropsychiatric CNVs; 465 controls carrying otherCNVs: 475 controls without large CNVs. In addition, 161 schizophrenia patients were recruited for the neuropsychological phenotyping.

Phenotyping. Encrypted identifiers of subjects were decrypted by a representative of the Icelandic Data Protection Authority and subjects were recruited to the study by a clinic overseen by the Icelandic Data Protection Authority. Psychologists and nurses phenotyping the participants were blind to genotype. Those working with the genetic data were blind to personal identifiers and could only work on the encrypted data set. Only a representative of the Data Protection Authority of Iceland holds the key for encrypting and decrypting the personal identifiers. Genotypes are only linked to encrypted identifiers. Approval for this study was obtained from the National Bioethics Committee of Iceland and the Icelandic Data Protection Authority. Written informed consent was obtained from all participants or their guardians before blood samples or phenotypic data were obtained. All sample identifiers were encrypted in accordance with the regulations of the Icelandic Data Protection Authority.

GAF[42] score was used to rate participants overall level of functioning and their ability to carry out activities of daily living. The scale was rated by the tester with respect to psychosocial, social, and occupational functioning. All participants were also interviewed using the MINI[5] edition 5.0.0. The MINI was designed as a brief structured interview for the major Axis I psychiatric disorders in ICD-10 and DSM-IV.

To assess cognitive function, logical memory I and II from the Wechsler Memory Scale (3rd edn) (WMS-III)[32] was used to assess memory. An Icelandic translation of the test was used (unstandardized). Two variables from the test were calculated; immediate memory that is the total item score immediately after the reading of story A and after both readings of story B, and delayed memory that is the total item score from both stories after 30 min delay. The average of the two scores was used in the analysis.

Verbal fluency was assessed using the COWAT[33] and the category naming test[34]; animal naming. In COWAT the subject is required to name as many words as he or she can that begin with a certain letter in one minute, the letters H and S were used in the Icelandic translation (unpublished). For the analysis a mean score of the number of words registered with each of the two letters were calculated (verbal fluency) and for category fluency the number of animals registered (category fluency).

The Stroop test[35] is a measure of selective attention and the ability to block out irrelevant stimuli. An Icelandic translation (unpublished), derived from the Golden version[43], was used in this study. In the first trial the participant is asked to read the names of colours written in black ink. In the second trial the participant has to name the colour of words written in coloured ink, and in the last (the main) trial the participant has to name the colour of the ink of a word which is actually the name of another colour. Two measures from the Stroop test were used in the analysis; the time it took to finish trial 1, and the interference score which is the difference in time to complete trial 3 minus the time for completing trial 2.

For visual scanning and mental flexibility, TMT A and B[36] were used administered. Trail A is a measure of psychomotor speed and attention and trail B is thought to be a test of flexibility of thinking. A measure of the time it took to finish trail A and a derived score of the time it took to finish trail B minus trail A was used in the analysis.

The WCST was designed to assess abstract reasoning and the ability to shift cognitive strategies in response to environmental cues[37]. A computerized version of the test was used[44]. The variable used in our analysis is per cent perseverative errors, which reflects the ratio of perseverative errors to the number of trials administered. Perseverative errors are made when participant persists in responding to the old rule after the rule has changed.

The SWM subtest from the CANTAB battery was used, which gives a measure of spatial working memory[38]. The measure used in this analysis was between-search errors for 6 boxes, which is a count of times the subject revisits a box where the token has previously been found. This is thought to rely on the long-term spatial memory system as the subject has to remember the location for some time and through interferences[45].

RVIP, a subtest from the CANTAB battery was used to access vigilance, which is the ability to sustain attention on one or more items over a period of time[39]. The main variable is A′, which is a signal detection measure of sensitivity to errors regardless of error tendency. This is a measure of the subjects ability to detect target sequences by using p(hit) and p(false). p(hit) is the probability of a hit; the proportion of correct responses that are given when a target sequence is presented on the screen. p(false) is the probability of a false alarm; the proportion of responses when there is no target sequence presented on the screen.

Intelligence quotient (IQ) was assessed using the WASI-I. The WASI-I test includes four subtests: vocabulary and similarities, both tests of verbal IQ, and matrix reasoning and block design, both tests of performance IQ[40].

The WASI-I test has been translated into Icelandic for standardization that is in progress and this study has been a part of that work. Here the healthy control group was used to make local norms by calculating the mean and s.d. for each of the age groups used in the US version of the WASI-I. *Z*-scores were then calculated for every subtest for each participant, and the mean of the subtests was transformed into an IQ score having a mean of 100 and a s.d. of 15 in the healthy control group.

A fraction of the participants were tested with an older translation of the WASI-I with two subtests, vocabulary and matrix reasoning. Thirty-nine subjects were tested using both editions of the test (with more than one year apart), and Pearson's correlation coefficient between the two measures in these subjects was 0.66. No significant effect of test type (new or old translation of WASI-I version) was found in any group included in the study.

The adult mathematical history questionnaire (AMHQ) described here is modelled after the adult reading history questionnaire (ARHQ)[12]. The AMHQ consists of six questions, each scored on a Likert-type scale ranging from 0 to 4.

The questions were: 1. Did you experience any difficulties in learning math in elementary school? 2. How was your math performance compared to your classmates in elementary school? 3. How do you rate your math skills now compared to people your age with a comparable education level? 4. Did you experience any difficulties learning the multiplication table in elementary school? 5. How much extra help did you need when learning math in elementary school? 6. What is your current attitude towards math?

The six questions were selected to assess the degree to which adults have experienced symptoms of specific disorder of arithmetical skills, or dyscalculia (F81.2), which according to the ICD-10 criteria[20], "Involves a specific impairment in arithmetical skills that is not solely explicable on the basis of general mental retardation or of inadequate schooling. The deficit concerns mastery of basic computational skills of addition, subtraction, multiplication, and division rather than of the more abstract mathematical skills involved in algebra, trigonometry, geometry, or calculus."

The score for the AMHQ scale ranges from 0 to 24 with higher scores indicating greater impairment. Internal consistency reliability ($\alpha = 0.90$) was assessed using Cronbach's α from AMHQ results of a large survey sample ($n = 2{,}757$). An exploratory factor analysis of this data set combining 28 items from ARHQ-Ice (22) and AMHQ (6) found that association with all three previously reported ARHQ subscale factors (dyslexic symptoms, current reading and memory)[11] were replicated, and all six AMHQ items had high factor loadings (≥ 0.55) on a separate fourth factor. This further confirms the internal consistency of the AMHQ scale and suggests an independence of the arithmetical disorder scale from the ARHQ total scale representing specific reading disorder or dyslexia and its three subscales[11]. Concurrent validity was assessed by comparing AMHQ scores of adults ($n = 39$) who had by formal neuropsychological evaluation been diagnosed as children with specific disorder of arithmetical skills (F81.2) and population controls without diagnosis of psychiatric or learning disorders and no learning disorder by self-report ($n = 564$). A significant difference in mean AMHQ scores was observed for these groups; that is, 17.8 (s.d. = 6.5) and 8.1 (s.d. = 5.3), respectively ($P < 0.001$).

For statistical analysis of cognitive traits, scores from each cognitive test or questionnaire were inverse normally transformed. They were then adjusted for sex, age at testing and, where indicated, IQ based on data from controls only. Final scores were shifted and scaled so that controls had a mean of 0 and a standard deviation of 1, and also arranged so that higher scores indicated greater impairment. To take the information on relatedness of the individuals into account, CNV carriers or schizophrenia patients were compared with controls using generalized least-squares regression with a variance–covariance matrix based on the kinship coefficient of each pair of individuals. Meiotic distance between neuropsychiatric CNV control carriers evaluated for cognitive traits can be found in Supplementary Table 4b.

The sample sizes obtained resulted in about 80% power to detect a difference of around 0.4 s.d. in the neuropsychiatric CNV control or schizophrenia versus population control comparisons, and about the same amount of power to detect a difference of around 0.3 s.d. for the other CNV control versus population control comparison. For the individual neuropsychiatric CNVs with the smallest sample size ($n = 5$), there was approximately 80% power to detect a difference of about 2.5 s.d.

Fecundity. deCODE genetics has built a nation-wide genealogy database for its genetic studies. The database contains information on year of birth and numbers of children of Icelanders. An encrypted version of the genealogy database was used for studying the fecundity in patient and CNV groups.

Mixed-effects Poisson generalized linear models (GLMs) were used to examine the association of fecundity with various neuropsychiatric CNVs and schizophrenia. The number of children at age 45 or older was regressed on sex, year of birth (included as factors for each 5 year birth cohort), sex-year of birth interaction (for each birth cohort factor), sibship (to account for relatedness) and the CNV or disorder of interest. All predictors were modelled as fixed effects except for sibship, which was random. A second set of models including a sex-CNV/disorder interaction term were also fit.

Neuroimaging. MRI examinations were conducted on a 1.5 T whole body Philips Achieva scanner. Scans were performed with a sagittal 3D fast T1-weighted gradient echo sequence (TR 8.6 ms, TE 4.0 ms, flip angle 8 degrees, slice thickness 1.2 mm, matrix 192 × 192, field of view 240 × 240 mm). Quality control of the MRI images consisted of a test of image homogeneity covariance and noise estimation (VBM8 toolbox; Gaser, http://dbm.neuro.uni-jena.de/author/admin/) as well as visual inspection.

For voxel-based morphometry, high-resolution T1-weighted images were processed according to the unified segmentation model[46] with SPM8 (statistical parametric mapping, Wellcome Department of Cognitive Neurology http://www.fil.ion.ucl.ac.uk/spm) and Matlab 8b software (The Mathworks). In brief, this method involves an iterated scheme of bias correction, segmentation into white matter, grey matter and cerebrospinal fluid and warping of prior images in stereotactic space to the data, which is repeated until no significant change occurs anymore. During normalization, images were interpolated to isotropic 1 × 1 × 1 mm voxels. The VBM8-toolbox extends this model with a partial volume estimation to account for partial volume effects and the application of a spatially adaptive non-local means (SANLM) filter[47] for bias correction. Normalization to stereotactic space consisted of a linear affine registration and a linear deformation corresponding to a high-dimensional DARTEL normalization[48] implemented in VBM8. The resulting probability maps were modulated, that is, intensity-corrected for local volume changes during normalization, to make them more sensitive to the distribution of grey matter and white matter volume. Modulation was limited to nonlinear warping; global differences in brain volume were thus excluded in the modulated probability maps. The modulated maps were smoothed with a 12-mm FWHM kernel.

For statistical analysis of neuroimaging data from MRI subjects carrying the 15q11.2(BP1-BP2) CNV, copy number effects at 15q11.2 (duplication > control > deletion and deletion > control > duplication) on regional brain volume were examined on a voxel-by-voxel basis with a multiple regression model using SPM8; age and gender were included as covariates of no interest.

An interaction between performance on neuropsychological tests that indicated a genetic dosage effect and copy number at 15q11.2, on regional grey matter volume was tested with a multiple regression analysis (SPM8); age and gender were included as covariates of no interest.

Effects on grey matter volume were reported as significant when whole-brain voxel-level FWE-corrected P value was less than 0.05. Additional region-of-interest (ROI) analyses were performed in the following regions found to show both functional and structural abnormalities in a recent meta-analysis of subjects with high risk of schizophrenia[21]: anterior cingulate and medial frontal cortex, and bilateral insula extending into temporal and parietal cortex. Results of these ROI analysis were considered significant at $P < 0.05$ voxel level, FWE-corrected.

ARTICLE

doi:10.1038/nature12867

Biochemical reconstitution of topological DNA binding by the cohesin ring

Yasuto Murayama[1] & Frank Uhlmann[1]

Cohesion between sister chromatids, mediated by the chromosomal cohesin complex, is a prerequisite for faithful chromosome segregation in mitosis. Cohesin also has vital roles in DNA repair and transcriptional regulation. The ring-shaped cohesin complex is thought to encircle sister DNA strands, but its molecular mechanism of action is poorly understood and the biochemical reconstitution of cohesin activity *in vitro* has remained an unattained goal. Here we reconstitute cohesin loading onto DNA using purified fission yeast cohesin and its loader complex, Mis4^{Scc2}–Ssl3^{Scc4} (*Schizosaccharomyces pombe* gene names appear throughout with their more commonly known *Saccharomyces cerevisiae* counterparts added in superscript). Incubation of cohesin with DNA leads to spontaneous topological loading, but this remains inefficient. The loader contacts cohesin at multiple sites around the ring circumference, including the hitherto enigmatic Psc3^{Scc3} subunit, and stimulates cohesin's ATPase, resulting in efficient topological loading. The *in vitro* reconstitution of cohesin loading onto DNA provides mechanistic insight into the initial steps of the establishment of sister chromatid cohesion and other chromosomal processes mediated by cohesin.

The cohesin complex is a central player in chromosome biology[1–5]. Defects in cohesin and its regulators are responsible for chromosome missegregation in human cancers and are the cause of Cornelia de Lange syndrome, a severe developmental disorder[6,7]. Despite notable advances[8–10], our molecular understanding of cohesin function remains vague. The cohesin complex consists of a dimer of structural maintenance of chromosomes subunits, Psm1^{Smc1} and Psm3^{Smc3}, long coiled coil proteins that interact at their hinge as well as their ABC-type ATPase head domains to form large proteinaceous rings[11,12]. The head interaction is stabilized by the kleisin subunit Rad21^{Scc1}. Several additional subunits associate with this ring assembly, including the essential Psc3^{Scc3} subunit, whose function remains poorly understood[12–16]. Cohesin is thought to promote sister chromatid cohesion by entrapping replicated sister chromatids within the ring's circumference[17], but how the Mis4^{Scc2}–Ssl3^{Scc4} cohesin loader[18], ATP hydrolysis by cohesin[19,20] and cohesion establishment during S phase[13,21,22] contribute to topological cohesin loading remains unknown. It also remains unknown whether the roles of cohesin outside of sister chromatid cohesion involve topological DNA binding.

The cohesin loader binds to DNA

We purified the fission yeast Mis4^{Scc2}–Ssl3^{Scc4} cohesin loader complex after overexpression of its two subunits in fission yeast (Fig. 1a and Extended Data Fig. 1a). We used a similar strategy to purify the large Mis4^{Scc2} subunit by itself. On the basis of its hydrodynamic properties, Mis4^{Scc2}–Ssl3^{Scc4} is a moderately elongated, heterodimeric protein complex (Fig. 1b and Extended Data Fig. 1b). Because Mis4^{Scc2} contains a putative leucine zipper, we investigated DNA binding of Mis4^{Scc2}–Ssl3^{Scc4}. We detected concentration-dependent DNA binding of Mis4^{Scc2}–Ssl3^{Scc4} to double-stranded DNA (dsDNA) (Fig. 1c, d and Extended Data Fig. 1c). Single-stranded DNA (ssDNA) was bound poorly, whereas a Y-fork DNA substrate, mimicking open DNA structures that might exist at some physiological cohesin loading sites, showed no increased affinity compared to dsDNA. The dsDNA preference over ssDNA was confirmed in a competition assay (Extended Data Fig. 1d). DNA binding was strongest at low salt concentrations, but remained detectable under physiological conditions (Extended Data Fig. 1e). The Mis4^{Scc2} subunit alone displayed DNA-binding properties indistinguishable from the Mis4^{Scc2}–Ssl3^{Scc4} complex (Fig. 1d and Extended Data Fig. 1d). These results suggest that the cohesin loader makes direct contact with dsDNA, which the Mis4^{Scc2} subunit is largely responsible for.

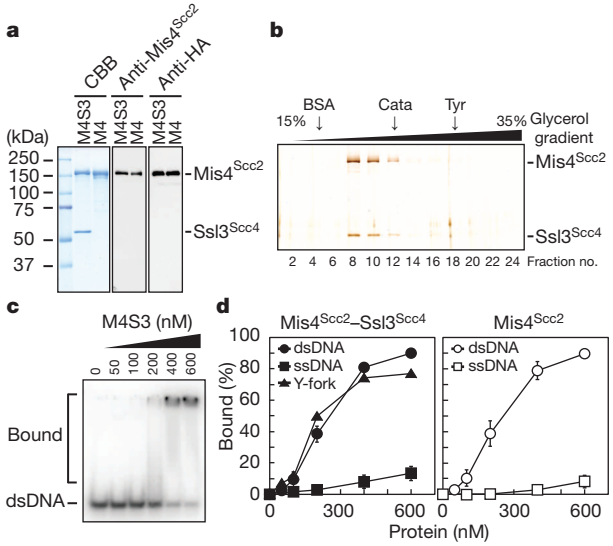


Figure 1 | Mis4^{Scc2}–Ssl3^{Scc4} is a DNA-binding protein. **a**, Purified Mis4^{Scc2}–Ssl3^{Scc4} complex (M4S3) and its Mis4^{Scc2} subunit (M4) were analysed by SDS–PAGE and Coomassie blue staining (CBB) or western blotting using antibodies directed against Mis4 or its carboxy-terminal haemagglutinin (HA) epitope. **b**, Glycerol gradient centrifugation of the Mis4^{Scc2}–Ssl3^{Scc4} complex, followed by SDS–PAGE and silver staining. Bovine serum albumin (BSA), catalase (Cata) and thyroglobulin (Tyr) were size markers. **c**, DNA binding of Mis4^{Scc2}–Ssl3^{Scc4}, using an electrophoretic mobility shift assay. **d**, Quantification of DNA binding by Mis4^{Scc2}–Ssl3^{Scc4} (left) and Mis4^{Scc2} (right). Mean and standard deviation from three independent experiments are shown.

[1]Chromosome Segregation Laboratory, Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK.

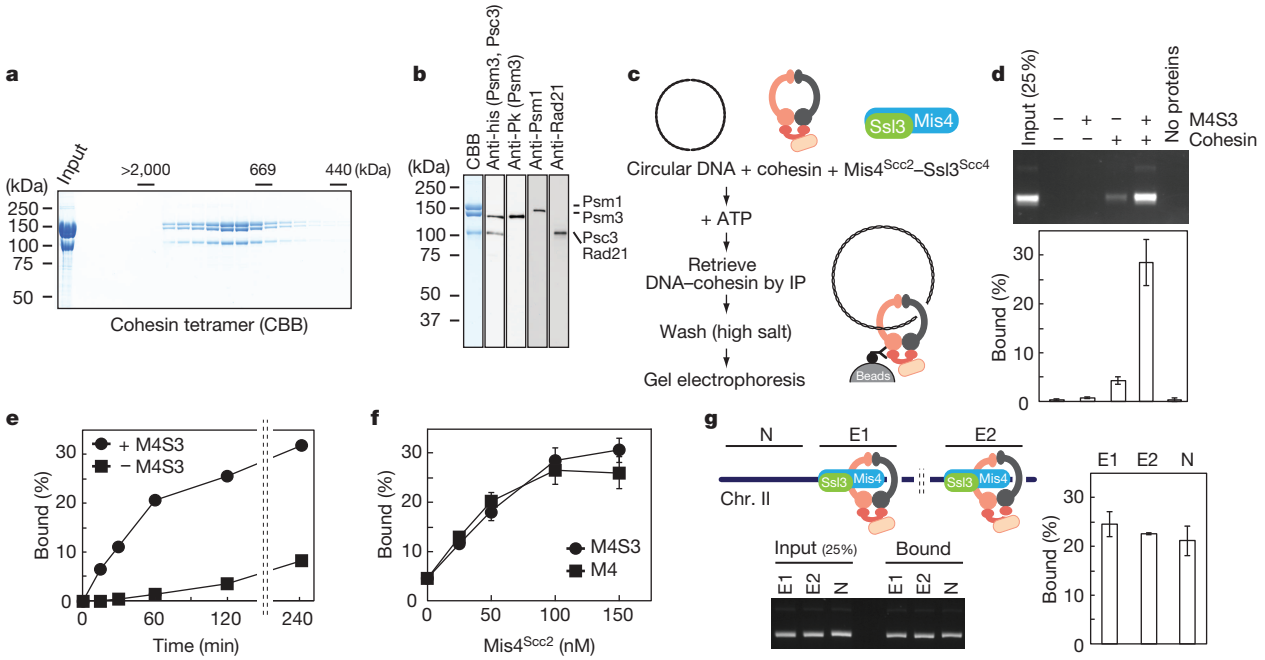


Figure 2 | *In vitro* reconstitution of cohesin loading onto DNA. **a**, Gel filtration of the cohesin complex. **b**, Purified cohesin, analysed by SDS–PAGE and Coomassie blue staining or western blotting as indicated. his, hexameric histidine; Pk, paramyxovirus SV5 Pk1 epitope. **c**, Schematic of the cohesin loading assay. **d**, Agarose gel electrophoresis and quantification of recovered DNA during the loading reaction. IP, immunoprecipitation. **e**, **f**, Time course with or without $Mis4^{Scc2}$–$Ssl3^{Scc4}$ (**e**) and titration of $Mis4^{Scc2}$–$Ssl3^{Scc4}$ or $Mis4^{Scc2}$ only (**f**). **g**, The loading reaction was carried out using rcDNA containing fission yeast cohesin-bound (E1, E2) or -unbound (N) sequences. Mean and standard deviation from at least three independent experiments are shown in **d**, **f** and **g**.

Topological cohesin loading *in vitro*

We purified the fission yeast cohesin complex after overexpression of its four essential subunits—Psm1, Psm3, Rad21 and Psc3—in budding yeast (Fig. 2a, b and Extended Data Fig. 2a). The complex showed ATP-independent dsDNA binding, as previously reported for cohesin[8,23], that was independent of the substrate topology but was salt sensitive (Extended Data Fig. 2b, c). Topological DNA binding by cohesin is expected to be salt resistant[10,17,18].

Taking these properties into account, we devised an assay to detect topological cohesin loading onto DNA (Fig. 2c). Cohesin and a relaxed circular DNA (rcDNA) substrate were mixed in the presence of ATP. After incubation, cohesin was immunoprecipitated. The cohesin beads were washed in high-salt buffer, and then DNA that remained bound was eluted and analysed by gel electrophoresis.

In the absence of protein, or in the presence of $Mis4^{Scc2}$–$Ssl3^{Scc4}$ only, no DNA was recovered (Fig. 2d). About 5% of input DNA was recovered when we performed the reaction in the presence of cohesin. The amount of bound DNA increased fivefold when $Mis4^{Scc2}$–$Ssl3^{Scc4}$ was also included. This suggests that the cohesin loader promotes salt-resistant cohesin loading onto DNA. The bulk of DNA binding occurred within 1 h, followed by a slower increase up to 4 h (Fig. 2e). DNA binding in the absence of $Mis4^{Scc2}$–$Ssl3^{Scc4}$ similarly increased over time, albeit at a lower level, indicating that it might also be the consequence of an active loading process. The salt-resistant loading onto DNA required that the incubation itself was performed at low salt concentrations, excluding the possibility of a high-salt artefact (Extended Data Fig. 3a). Both $Mis4^{Scc2}$ and the $Mis4^{Scc2}$–$Ssl3^{Scc4}$ complex caused indistinguishable concentration-dependent stimulation of cohesin loading (Fig. 2f), suggesting that the ability to load cohesin onto DNA is contained within the $Mis4^{Scc2}$ subunit of the cohesin loader.

Cohesin and its loader localize to specific chromosomal loci[24,25]. We modified our DNA substrate to contain fission yeast DNA sequences that are *in vivo* enriched for, or free from, $Mis4^{Scc2}$–$Ssl3^{Scc4}$ and cohesin[25]. No differences were observed when comparing cohesin loading using these substrates, indicating that DNA sequence does not affect cohesin loading, at least under our *in vitro* conditions (Fig. 2g).

As expected, different closed circular DNAs served as efficient substrates in the loading reaction, but not linear DNA (Extended Data Fig. 3b). To confirm that the interaction was topological, we retrieved cohesin from a loading reaction, then linearized the bound circular DNA. This released cleaved DNA into the supernatant, whereas residual uncleaved DNA remained cohesin-bound on the beads (Fig. 3a, b). Topological loading was also observed in a loading reaction containing the $Mis4^{Scc2}$ subunit only (Extended Data Fig. 3c). The lower level of DNA loaded onto cohesin in the absence of any loader was also released by linearization (Fig. 3b), suggesting that cohesin achieves topological loading onto DNA independently of a cohesin loader, albeit inefficiently.

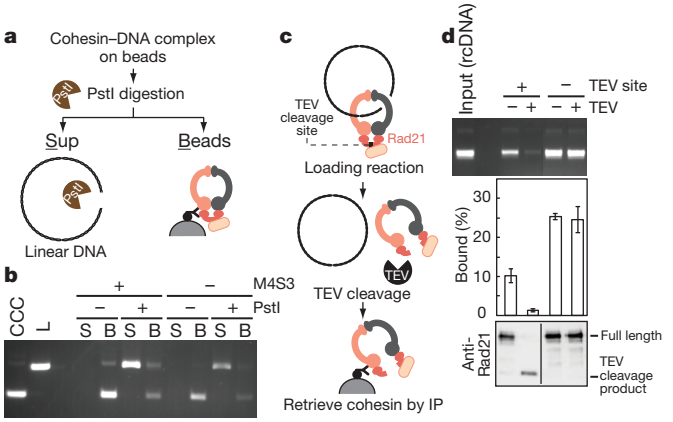


Figure 3 | Topology-mediated DNA binding of cohesin. **a**, Schematic of the DNA-release experiment. **b**, Gel image of the experiment. Covalently closed circular (CCC) input DNA and linear form (L), supernatant (S) and bead (B) fractions of experiments in the presence or absence of $Mis4^{Scc2}$–$Ssl3^{Scc4}$, with or without PstI digestion, are shown. **c**, Diagram of the DNA-release experiment by TEV protease cleavage of Rad21. **d**, Agarose gel of recovered rcDNA and its quantification, the mean and standard deviation of three independent experiments, are shown. Rad21 cleavage was monitored by western blotting.

Cohesin dissociates from chromosomes to trigger anaphase following cleavage of its kleisin subunit by the protease separase[14,26]. To test whether our *in vitro*-loaded cohesin recapitulates this, we replaced one of the two separase-recognition sequences in the kleisin subunit Rad21 with a tobacco etch virus (TEV) protease-recognition motif (Fig. 3c). After loading this modified cohesin complex, we added TEV protease. This led to efficient Rad21 cleavage and concomitant loss of DNA from TEV-cleavable cohesin, but not from cohesin lacking the TEV-recognition site (Fig. 3d). Thus, cohesin achieves its final, cleavage-sensitive topological state on DNA from its initial loading. We conclude that topological loading onto DNA is an intrinsic activity of the cohesin complex, although it is greatly facilitated by the Mis4^{Scc2}–Ssl3^{Scc4} cohesin loader complex, particularly the Mis4^{Scc2} subunit.

The loader engages cohesin's ATPase

To investigate the contribution of cohesin's ATPase, we carried out loading reactions in the absence or presence of ATP, ADP or non-hydrolysable ATP analogues and found that ATP hydrolysis is required for cohesin loading onto DNA (Fig. 4a). Cohesin complexes harbouring ATPase-inactivating point mutations in the Walker A ATPase motif of Psm1 or Psm3 also failed to be loaded onto DNA by Mis4^{Scc2}–Ssl3^{Scc4} *in vitro* (Fig. 4b and Extended Data Fig. 2d). Therefore the ATP-hydrolysis-dependent DNA-binding step that has been inferred from mutant analysis *in vivo*[19,20] is most likely that of topological loading onto DNA.

Fission yeast cohesin hydrolysed ATP at a rate of 3.7 min^{-1}, dependent on its intact ATPase (Fig. 4c, d). The addition of DNA or Mis4^{Scc2}–Ssl3^{Scc4} did not change this rate, but the addition of both led to a marked increase in ATP hydrolysis. This suggests that Mis4^{Scc2}–Ssl3^{Scc4} stimulates cohesin loading once the loader, cohesin and DNA come together. The Mis4^{Scc2} subunit alone had a similar effect to the loader complex. Linear DNA was as efficient as circular DNA in stimulating ATP hydrolysis (Fig. 4d), suggesting that DNA topology might not affect cohesin loading, but rather the retention of cohesin on DNA once loaded.

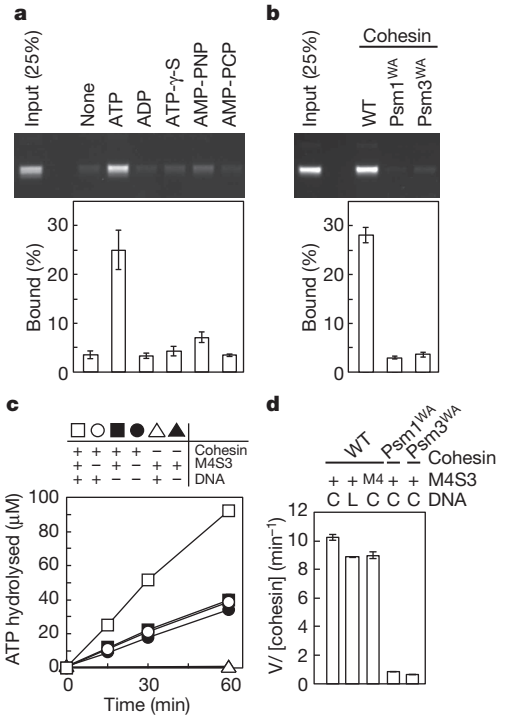


Figure 4 | Mis4^{Scc2}–Ssl3^{Scc4} stimulates cohesin's ATPase. **a**, **b**, Loading reactions in the presence or absence of ATP or its derivatives (**a**), or comparing wild-type or Walker A (WA) motif mutant cohesin (**b**). **c**, Time-course analysis of ATP hydrolysis by cohesin with or without Mis4^{Scc2}–Ssl3^{Scc4} or rcDNA. **d**, ATP-hydrolysis rates of wild-type and mutant cohesin derived from similar time-course analyses. rcDNA (C) or linear DNA (L) and a reaction with Mis4^{Scc2} only was included. The mean and standard deviation of three independent experiments are shown.

Multiple interactions around the ring

We next analysed the physical interaction between Mis4^{Scc2}–Ssl3^{Scc4} and cohesin. Consistent with previous observations[10,20], the two purified complexes co-immunoprecipitated, suggesting that they directly interact. Mis4^{Scc2}–Ssl3^{Scc4} (or Mis4^{Scc2} by itself) also interacted with the purified Psm1–Psm3 dimer, with the Psm3 subunit alone, as well as with the purified Psc3 subunit (Extended Data Fig. 4). To map the interacting regions, we used tiling peptide arrays covering each of the four cohesin subunits. They were probed with Mis4^{Scc2}–Ssl3^{Scc4} that was subsequently detected by immunoblotting. As an example, Fig. 5a shows the identification of a candidate Mis4^{Scc2}–Ssl3^{Scc4} interaction site on Psm1, close to its hinge. We identified putative loader interaction sites on all four cohesin subunits, located at positions along the cohesin ring circumference (Fig. 5b and Extended Data Fig. 5). The contact regions within Psm1, Psm3 and Psc3 show evolutionary conservation. In particular, two of the interacting regions in Psc3 delineate the 'stromalin homology domain'[27] that is characteristic of Psc3 orthologues (Extended Data Fig. 6a–c).

A second set of peptide arrays was used to identify amino acids critical for these interactions. To verify the importance of the key residues identified (Extended Data Fig. 6d–f), we made corresponding point mutations in the *psm1*$^+$, *psm3*$^+$ and *psc3*$^+$ genes and ectopically expressed them in fission yeast cells carrying temperature-sensitive mutations in the respective subunits (Extended Data Fig. 7). Point mutations in Psm3 and each of the three interaction sites in Psc3 caused failure to complement temperature sensitivity. In the case of Psm1, we observed sensitivity to DNA-damaging agents, a known characteristic of reduced cohesin function[28]. Thus, Mis4^{Scc2}–Ssl3^{Scc4} interacts with cohesin at multiple sites around its ring circumference, at places that are important for cohesin function.

Ring contacts promote cohesin loading

The interaction of Mis4^{Scc2}–Ssl3^{Scc4} with Psc3 was of particular interest as no molecular role has yet been ascribed to this essential cohesin subunit[12–16]. The interaction between the two components was noticeably reduced by point mutations in Psc3, particularly those within the stromalin homology domain (Fig. 5c). When expressed in fission yeast, all of the interaction site mutants failed to rescue the sister chromatid cohesion defect of a *psc3-303* temperature-sensitive strain (Fig. 5d), confirming that interaction of Psc3 with the cohesin loader is important for sister chromatid cohesion.

To define the role of Psc3 in cohesin loading, we carried out loading reactions with a cohesin trimer including Psm1, Psm3 and Rad21, and lacking Psc3. The cohesin trimer by itself did not detectably bind rcDNA, although addition of Mis4^{Scc2}–Ssl3^{Scc4} resulted in a small degree of loading (Fig. 5e and Extended Data Fig. 8a). Adding back Psc3 to the cohesin trimer also gave a low level of DNA loading. However, Psc3 together with Mis4^{Scc2}–Ssl3^{Scc4} substantially restored cohesin loading. Likewise, DNA-dependent ATPase activity of the cohesin trimer was low, but was coordinately stimulated by addition of Psc3 and Mis4^{Scc2}–Ssl3^{Scc4} (Extended Data Fig. 8b). The rescue of cohesin loading by Psc3 was reduced by point mutations in loader interaction sites, especially those in the stromalin homology domain (Fig. 5e). This effect was augmented at a higher incubation temperature of 37 °C, when small interaction defects might be less well tolerated. Each of the loader interaction-deficient Psc3 proteins stimulated the loader-independent ATPase of the cohesin trimer to a similar extent as wild-type Psc3 (Extended Data Fig. 8c), suggesting that they retain functional interactions within the cohesin complex. Thus, a specific interaction of Psc3 with Mis4^{Scc2}–Ssl3^{Scc4} is required to promote cohesin loading.

We also investigated whether interaction of Mis4^{Scc2}–Ssl3^{Scc4} with Psm1 and Psm3 contributes to cohesin loading. We introduced loader interaction site mutations into both Psm1 and Psm3. This did not affect

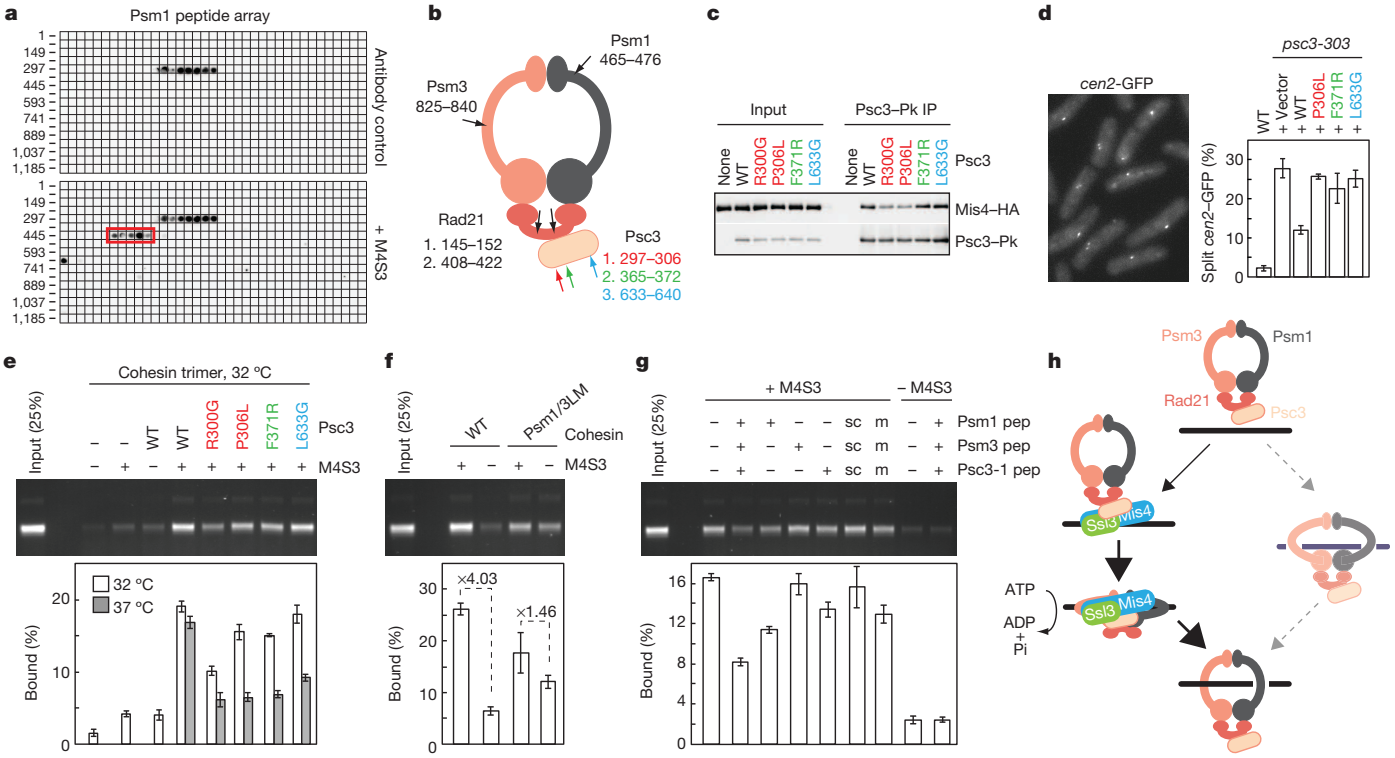


Figure 5 | Mis4^{Scc2}–Ssl3^{Scc4} interactions around the cohesin ring. **a**, Identification of a Mis4^{Scc2}–Ssl3^{Scc4} interaction site on a Psm1 tiling peptide array. Starting positions of the first peptide in every other row are indicated. **b**, Summary of Mis4^{Scc2}–Ssl3^{Scc4} interaction sites on the four cohesin subunits **c**, Co-immunoprecipitation of wild-type (WT) and interaction site mutant Psc3-Pk with Mis4^{Scc2}–Ssl3^{Scc4}. **d**, Rescue of sister chromatid cohesion after *psc3-303* inactivation by wild-type or interaction site mutant Psc3. 100 cells were scored each in three independent experiments; means and standard deviations are shown. **e**, Loading reactions of a cohesin trimer were supplemented with wild-type or interaction site mutant Psc3. **f**, Wild-type and loader interaction site mutation (Psm1/3^{LM}) cohesin complexes were compared. **g**, Wild-type (+), scrambled (sc) or point mutant (m) peptides corresponding to Mis4^{Scc2}–Ssl3^{Scc4} interactions sites were included in loading reactions. Panels **e**–**g** present means and standard deviations from three independent experiments. **h**, Model for cohesin loading onto DNA.

cohesin complex stability or its loader-independent association with DNA (Fig. 5f and Extended Data Fig. 2d). However, loading was stimulated less than 1.5-fold by Mis4^{Scc2}–Ssl3^{Scc4}, whereas DNA loading of wild-type cohesin complexes is stimulated over fourfold by the loader. This suggests that contacts of the cohesin loader with Psm1 and Psm3 also contribute to cohesin loading.

In a complementary approach, we observed competition of the loading reaction by peptides corresponding to Mis4^{Scc2}–Ssl3^{Scc4} interaction sites on Psm1, Psm3 and Psc3. *In vitro* cohesin loading onto DNA was reduced by addition of each individual peptide, an effect that was augmented by combining the three peptides (Fig. 5g). Similar addition of peptides with a scrambled sequence, or containing point mutations that reduce the Mis4^{Scc2}–Ssl3^{Scc4} interaction, did not elicit inhibition. Peptide addition did not affect Mis4^{Scc2}–Ssl3^{Scc4}-independent DNA loading of cohesin, indicating that the interaction site peptides target the cohesin loader. Together, our results suggest that multiple interactions of the cohesin loader along the circumference of the cohesin ring jointly promote cohesin loading onto DNA.

Discussion

We have reconstituted topological loading of cohesin onto DNA using purified fission yeast proteins. The biochemical characterization of this reaction has led to several surprises. Cohesin topologically binds to DNA independently of a loading factor or of cohesion establishment reactions. We believe that our experiments using purified components provide final proof for the idea that cohesin entraps DNA[12], which has been previously supported by evidence from less-well-defined and therefore less-conclusive systems[17]. A second structural maintenance of chromosomes (SMC) family ring complex, condensin, is thought to associate with budding yeast chromatin by topological embrace[29]. Condensin binding to chromatin may not, or only in part, depend on Mis4^{Scc2}–Ssl3^{Scc4} in fission and budding yeast[30,31]. In the case of prokaryotic SMC complexes, which show considerable similarity to their eukaryotic counterparts[32], a loader similar to the Mis4^{Scc2}–Ssl3^{Scc4} complex is unknown. Assuming that bacterial SMC complexes also embrace DNA, our results explain how SMC complexes can topologically bind DNA without the need for a loader. The intrinsic ability to entrap DNA also makes it possible that cohesin acts in DNA repair and transcriptional regulation by topological embrace.

Cohesin loading onto DNA *in vitro* was greatly facilitated by the cohesin loader Mis4^{Scc2}–Ssl3^{Scc4}, an essential protein complex in all eukaryotes studied. Unexpectedly, the Mis4^{Scc2} subunit alone harboured all activities required for the cohesin loading reaction. These findings contrast with the essential nature of Ssl3^{Scc4} (refs 18, 33). Budding yeast Ssl3^{Scc4} is required for protein stability of the Mis4^{Scc2} subunit *in vivo*[18]. In addition, Ssl3^{Scc4} could have a role in cohesin loading *in vivo* that was not recapitulated in our *in vitro* loading reaction. Loading occurred in a sequence-nonspecific fashion *in vitro*, whereas the cohesin loader occupies discrete locations on chromosomes *in vivo*[24,25]. In *Xenopus* egg extract, the cohesin loader is recruited to chromatin via an interaction with pre-replicative complexes that requires its Ssl3^{Scc4} subunit[34]. Ssl3^{Scc4} could mediate recruitment of the cohesin loader in the context of a chromatinized DNA substrate to ensure that cohesin loading occurs with spatial and temporal precision.

Mis4^{Scc2}–Ssl3^{Scc4} makes numerous contacts with the cohesin complex, including a prominent interaction with the previously enigmatic Psc3 subunit. These jointly facilitate cohesin loading onto DNA. By engaging into multiple contacts, Mis4^{Scc2}–Ssl3^{Scc4} could stabilize a transient conformation of the cohesin complex during the DNA-loading reaction that is otherwise energetically unfavourable (Fig. 5h). Recent

studies of the SMC-related DNA repair protein Rad50 have suggested how an ATP-hydrolysis-dependent conformational change of the ATPase domains could be transmitted along their coiled coil[35–37]. A similar ATP-dependent conformational change has also been implicated in ring opening of a mismatch repair ABC-type ATPase[38]. Given the large dimensions of the cohesin ring, $Mis4^{Scc2}$–$Ssl3^{Scc4}$ may function as a molecular 'shaft' to help transmit an ATP-dependent conformational change from the head domains along the coiled coil to the hinge[39]. The Psc3 subunit in this model would act as a molecular 'hitch' to connect the ATPase 'engine' to the $Mis4^{Scc2}$–$Ssl3^{Scc4}$ 'shaft'. This scenario offers a molecular explanation for how ATP hydrolysis at the cohesin heads leads to ring opening at the opposite side of the cohesin ring. The 'shaft' at the same time controls the ATPase engine to ensure it is only started when all components of the loading reaction are in place.

METHODS SUMMARY

Protein purification. The fission yeast $Mis4^{Scc2}$–$Ssl3^{Scc4}$ and cohesin complexes were overexpressed in fission and budding yeast cells, respectively, and purified using protein A affinity, heparin and size-exclusion chromatography. Psc3 was expressed and purified from *Escherichia coli*.

***In vitro* cohesin loading assay.** $Mis4^{Scc2}$–$Ssl3^{Scc4}$ and cohesin were incubated with a rcDNA substrate in the presence of ATP at 32 °C for 1 h. Cohesin was retrieved by immunoprecipitation and, after high-salt washes, cohesin-bound DNA was analysed by agarose gel electrophoresis.

Other assays. Electrophoretic mobility shift assays were carried out using ^{32}P-labelled 50-nucleotide oligonucleotides as DNA substrates and were analysed by polyacrylamide gel electrophoresis. ATPase activity was determined by monitoring the hydrolysis of [γ-^{32}P]-ATP by thin-layer chromatography. Protein interactions were analysed by co-immunoprecipitation of purified $Mis4^{Scc2}$–$Ssl3^{Scc4}$ and cohesin, or by probing peptide tiling arrays of the cohesin subunits, synthesized on cellulose membranes, with $Mis4^{Scc2}$–$Ssl3^{Scc4}$ followed by its immunodetection.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 May; accepted 13 November 2013.
Published online 1 December 2013.

1. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91,** 35–45 (1997).
2. Guacci, V., Koshland, D. & Strunnikov, A. A direct link between sister chromatid cohesion and chromosome condensation revealed through analysis of *MCD1* in *S. cerevisiae*. *Cell* **91,** 47–57 (1997).
3. Losada, A., Hirano, M. & Hirano, T. Identification of *Xenopus* SMC protein complexes required for sister chromatid cohesion. *Genes Dev.* **12,** 1986–1997 (1998).
4. Sjögren, C. & Nasmyth, K. Sister chromatid cohesion is required for postreplicative double-strand break repair in *Saccharomyces cerevisiae*. *Curr. Biol.* **11,** 991–995 (2001).
5. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451,** 796–801 (2008).
6. Musio, A. *et al.* X-linked Cornelia de Lange syndrome owing to *SMC1L1* mutations. *Nature Genet.* **38,** 528–530 (2006).
7. Solomon, D. A. *et al.* Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science* **333,** 1039–1043 (2011).
8. Losada, A. & Hirano, T. Intermolecular DNA interactions stimulated by the cohesin complex *in vitro*: implications for sister chromatid cohesion. *Curr. Biol.* **11,** 268–272 (2001).
9. Onn, I. & Koshland, D. *In vitro* assembly of physiological cohesin/DNA complexes. *Proc. Natl Acad. Sci. USA* **108,** 12198–12205 (2011).
10. Bermudez, V. P. *et al.* *In vitro* loading of human cohesin on DNA by the human Scc2-Scc4 loader complex. *Proc. Natl Acad. Sci. USA* **109,** 9366–9371 (2012).
11. Anderson, D. E., Losada, A., Erickson, H. P. & Hirano, T. Condensin and cohesin display different arm conformations with characteristic hinge angles. *J. Cell Biol.* **156,** 419–424 (2002).
12. Haering, C. H., Löwe, J., Hochwagen, A. & Nasmyth, K. Molecular architecture of SMC proteins and the yeast cohesin complex. *Mol. Cell* **9,** 773–788 (2002).
13. Tóth, A. *et al.* Yeast Cohesin complex requires a conserved protein, Eco1p (Ctf7), to establish cohesion between sister chromatids during DNA replication. *Genes Dev.* **13,** 320–333 (1999).
14. Tomonaga, T. *et al.* Characterization of fission yeast cohesin: essential anaphase proteolysis of Rad21 phosphorylated in the S phase. *Genes Dev.* **14,** 2757–2770 (2000).
15. Losada, A., Yokochi, T., Kobayashi, R. & Hirano, T. Identification and characterization of SA/Scc3p subunits in the *Xenopus* and human cohesin complexes. *J. Cell Biol.* **150,** 405–416 (2000).
16. Sumara, I., Vorlaufer, E., Gieffers, C., Peters, B. H. & Peters, J.-M. Characterization of vertebrate cohesin complexes and their regulation in prophase. *J. Cell Biol.* **151,** 749–762 (2000).
17. Haering, C. H., Farcas, A. M., Arumugam, P., Metson, J. & Nasmyth, K. The cohesin ring concatenates sister DNA molecules. *Nature* **454,** 297–301 (2008).
18. Ciosk, R. *et al.* Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins. *Mol. Cell* **5,** 243–254 (2000).
19. Weitzer, S., Lehane, C. & Uhlmann, F. A model for ATP hydrolysis-dependent binding of cohesin to DNA. *Curr. Biol.* **13,** 1930–1940 (2003).
20. Arumugam, P. *et al.* ATP hydrolysis is required for cohesin's association with chromosomes. *Curr. Biol.* **13,** 1941–1953 (2003).
21. Rolef Ben-Shahar, T. *et al.* Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science* **321,** 563–566 (2008).
22. Ünal, E. *et al.* A molecular determinant for the establishment of sister chromatid cohesion. *Science* **321,** 566–569 (2008).
23. Sakai, A., Hizume, K., Sutani, T., Takeyasu, K. & Yanagida, M. Condensin but not cohesin SMC heterodimer induces DNA reannealing through protein–protein assembly. *EMBO J.* **22,** 2764–2775 (2003).
24. Lengronne, A. *et al.* Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **430,** 573–578 (2004).
25. Schmidt, C. K., Brookes, N. & Uhlmann, F. Conserved features of cohesin binding along fission yeast chromosomes. *Genome Biol.* **10,** R52 (2009).
26. Uhlmann, F., Wernic, D., Poupart, M.-A., Koonin, E. V. & Nasmyth, K. Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. *Cell* **103,** 375–386 (2000).
27. Pezzi, N. *et al.* *STAG3*, a novel gene encoding a protein involved in meiotic chromosome pairing and location of *STAG3*-related genes flanking the Williams-Beuren syndrome deletion. *FASEB J.* **14,** 581–592 (2000).
28. Birkenbihl, R. P. & Subramani, S. Cloning and characterization of *rad21* an essential gene of *Schizosaccharomyces pombe* involved in DNA double-strand-break repair. *Nucleic Acids Res.* **20,** 6605–6611 (1992).
29. Cuylen, S., Metz, J. & Haering, C. H. Condensin structures chromosomal DNA through topological links. *Nature Struct. Mol. Biol.* **18,** 894–901 (2011).
30. D'Ambrosio, C. *et al.* Identification of *cis*-acting sites for condensin loading onto budding yeast chromosomes. *Genes Dev.* **22,** 2215–2227 (2008).
31. Furuya, K., Takahashi, K. & Yanagida, M. Faithful anaphase is ensured by Mis4, a sister chromatid cohesion molecule required in S phase and not destroyed in G_1 phase. *Genes Dev.* **12,** 3408–3418 (1998).
32. Gruber, S. & Errington, J. Recruitment of condensin to replication origin regions by ParB/SpoOJ promotes chromosome segregation in *B. subtilis*. *Cell* **137,** 685–696 (2009).
33. Bernard, P. *et al.* A screen for cohesion mutants uncovers Ssl3, the fission yeast counterpart of the cohesin loading factor Scc4. *Curr. Biol.* **16,** 875–881 (2006).
34. Takahashi, T. S., Basu, A., Bermudez, V., Hurwitz, J. & Walter, J. C. Cdc7–Drf1 kinase links chromosome cohesion to the initiation of DNA replication in *Xenopus* egg extracts. *Genes Dev.* **22,** 1894–1905 (2008).
35. Lammens, K. *et al.* The Mre11:Rad50 structure shows an ATP-dependent molecular clamp in DNA double-strand break repair. *Cell* **145,** 54–66 (2011).
36. Williams, G. J. *et al.* ABC ATPase signature helices in Rad50 link nucleotide state to Mre11 interface for DNA repair. *Nature Struct. Mol. Biol.* **18,** 423–431 (2011).
37. Lim, H. S., Kim, J. S., Park, Y. B., Gwon, G. H. & Cho, Y. Crystal structure of the Mre11–Rad50–ATPγS complex: understanding the interplay between Mre11 and Rad50. *Genes Dev.* **25,** 1091–1104 (2011).
38. Warren, J. J. *et al.* Structure of the human MutSα DNA lesion recognition complex. *Mol. Cell* **26,** 579–592 (2007).
39. Nasmyth, K. Cohesin: a catenase with separate entry and exit gates? *Nature Cell Biol.* **13,** 1170–1177 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to N. O'Reilly for peptide synthesis, A. Alidoust and N. Patel for fermentation and J. Hurwitz, T. Toda and members of the Chromosome Segregation Laboratory for discussion and comments on the manuscript. This work was supported by the European Research Council. Y.M. was supported by the Japanese Society for the Promotion of Science (JSPS).

Author Contributions Y.M. designed the study, performed all the experiments, analysed data and wrote the manuscript. F.U. designed and supervised the study and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.U. (frank.uhlmann@cancer.org.uk).

METHODS

Expression and purification of Mis4^{Scc2}–Ssl3^{Scc4}. The *mis4*$^{+}$ complementary DNA was amplified by PCR from a fission yeast *S. pombe* meiotic cDNA library (Yeast National BioResource Project, Osaka, Japan) and cloned under control of the *nmt1*$^{+}$ promoter into plasmid pREP1 (*LEU2*). An HA epitope and a protein A tag, separated by a PreScission protease recognition sequence, were added at the C terminus, yielding plasmid pMis4PA. The *ssl3*$^{+}$ cDNA was similarly cloned under control of the *nmt1*$^{+}$ promoter, without epitope tags, into pREP2 (*ura4*$^{+}$) generating pSsl3. Both pMis4PA and pSsl3 were together introduced into fission yeast (*h− ura4-D18 leu1-32*) and protein expression induced by growing the cells in Edinburgh minimal medium (EMM2) lacking thiamine at 30 °C for 15 h. Cell pellets were re-suspended in an equal volume of CLR buffer (50 mM Tris/HCl, pH 7.5, 1 mM dithiothreitol (DTT), 250 mM NaCl, 2.5 mM $MgCl_2$, 5 mM EGTA, 20% (v/v) glycerol, 0.2 mM sodium vanadate, 0.5 mM phenylmethylsulphonyl fluoride (PMSF) and a protease inhibitor cocktail (Roche)), frozen in liquid nitrogen and then broken in a freezer mill (SPEX CertiPrep 6850). The cell powder was thawed at 4 °C, then twice the volume of CLR buffer was added. The lysate was clarified at 4 °C by centrifugation at 45,000*g* for 30 min, and then at 200,000*g* for 1 h. The clarified lysate was mixed with IgG sepharose (GE Healthcare, 1 ml resin slurry per 50 ml of lysate) and RNase A (10 μg ml^{-1} final) at 4 °C for 3 h. The resin was washed with 15 bed volumes of R buffer (20 mM Tris/HCl, pH 7.5, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), 10% (v/v) glycerol) containing 250 mM NaCl and 0.5 mM PMSF, and then with 15 bed volumes of the same buffer lacking PMSF. The resin was then suspended in two bed volumes of R buffer containing 250 mM NaCl, 10 μg ml^{-1} RNase A and 5 U ml^{-1} of PreScission protease (GE Healthcare) and incubated overnight at 4 °C. For elution, 1.5 volumes of R buffer were added to adjust the salt concentration to 100 mM NaCl and the eluate was loaded onto a HiTrap Heparin HP column (GE Healthcare). The column was developed with a linear gradient from 100 mM to 1 M NaCl in R buffer. The peak fractions were pooled and loaded onto a Superdex 200 10/300 GL gel filtration column (GE Healthcare) that was developed in R buffer containing 150 mM NaCl. The peak fractions containing Mis4–Ssl3 were concentrated to 500 μl by centrifugal ultrafiltration (Amicon Ultra, Millipore). Typically, 0.5 mg of Mis4–Ssl3 was obtained from 30 g of fission yeast cells.

The Mis4 subunit was overexpressed and purified from fission yeast (*h− ura4-D18 leu1-32*), bearing pMis4PA only, following essentially the same procedure. Typically, 0.3 mg of Mis4 was obtained from 10 g of cells.

Expression and purification of cohesin. The Psm3-encoding cDNA, fused to 3 × Pk epitope and 7 × histidine tag at the C terminus, was cloned under control of the bidirectional budding yeast *S. cerevisiae GAL1–GAL10* promoter in the *GAL1* direction into the shuttle vector YIplac211 (*URA3*)[40]. The cDNA encoding Psm1 was cloned into the same plasmid in the direction of the *GAL10* promoter, yielding the plasmid YIpPsm1-Psm3. The *rad21*$^{+}$ cDNA, fused to an HA epitope and a protein A tag at the C terminus, separated by a PreScission protease recognition sequence, was cloned under control of the *GAL1* promoter into YIplac128 (*LEU2*). The *psc3*$^{+}$ cDNA, fused to a 7 × histidine tag at the amino terminus, was added in the direction of the *GAL10* promoter yielding the plasmid YIpRad21-Psc3. The linearized YIpPsm1-Psm3 and YIpRad21-Psc3 plasmids were sequentially integrated into budding yeast (*MAT***a**, *ade2-1*, *trp1-1*, *can1-100*, *leu2,3,112*, *his3-11,15*, *ura3-52*, *pep4Δ::HIS3MX*) at the *URA3* and *LEU2* loci, respectively. The resultant Psm1/Psm3/Rad21/Psc3-expressing cells were grown in YP medium containing 2% raffinose to an OD_{600} = 2.0 at 30 °C. Galactose was added to the culture at a final concentration of 2% and the cells were grown for further 4.5 h at 30 °C. Cell pellets after centrifugation were re-suspended in an equal volume of CLH buffer (50 mM HEPES/KOH, pH 7.5, 1 mM DTT, 300 mM NaCl, 20% (v/v) glycerol, 0.5 mM PMSF and protease inhibitor cocktail), frozen in liquid nitrogen and broken in a freezer mill. Cell lysate preparation and purification on IgG sepharose followed the procedure for purification of Mis4–Ssl3, but CLH buffer and H buffer (25 mM HEPES/KOH, pH 7.5, 0.5 mM TCEP, 10% (v/v) glycerol) containing 300 mM NaCl were used. Two volumes of R buffer were added to the eluate to bring the salt concentration to 100 mM before loading onto a HiTrap Heparin HP column. Bound proteins were eluted with steps of 300 mM, 600 mM and 1 M NaCl in R buffer. Cohesin was retrieved in the 600 mM NaCl fraction. This fraction was applied to a Superose 6 10/300 GL gel filtration column (GE Healthcare) that was developed in R buffer containing 200 mM NaCl. The peak fractions were concentrated to 400 μl by ultrafiltration. Typically, 0.5 mg of the cohesin complex was obtained from 30 g of budding yeast cells. Cohesin containing a Rad21 TEV cleavage site, cohesin bearing K38I mutations in the Walker A ATPase motif of Psm1 (Psm1^{WA}) or Psm3 (Psm3^{WA}) and the cohesin trimer lacking Psc3 were purified using the same procedure. For expression of the cohesin Psm1/3 loader interaction mutant (Psm1/3^{LM}, containing Psm1(K467E, R468E) and Psm3(R828E, R829E)), an additional plasmid expressing the Psm3 mutant under control of the *GAL1* promoter, based on YIplac204, was integrated at the *TRP1* locus. This plasmid also contained the budding yeast *GAL4* gene under control of the *GAL10* promoter to improve galactose-induced protein expression. Cohesin Psm1/3^{LM} was purified as above.

Expression and purification of the Psm1–Psm3 heterodimer and of Psm3. For purification of the Psm1–Psm3 heterodimer, the *psm1*$^{+}$ cDNA was cloned into pREP1, yielding pPsm1. The *psm3*$^{+}$ cDNA, fused to a 3 × Pk and 7 × histidine tag at the C terminus, was cloned into pREP2 to yield pPsm3PkH. Both proteins were simultaneously overexpressed in fission yeast and cell lysate was prepared essentially as described for Mis4–Ssl3 purification, using CLH buffer. The clarified lysate was mixed with 1 ml nickel-nitrilotriacetic acid (Ni-NTA) agarose slurry (Qiagen) per 50 ml of lysate for 4 h at 4 °C. The beads were washed with 30 volumes of H buffer containing 300 mM NaCl and the bound proteins were eluted with H buffer containing 300 mM NaCl and 250 mM imidazole. The eluate was loaded onto a Superose 6 10/300 GL gel filtration column that was developed in R buffer containing 200 mM NaCl. The peak fractions were concentrated by ultrafiltration.

Psm3 was expressed in fission yeast cells harbouring the pPsm3PkH plasmid only. Purification followed the same protocol as described for the Psm1–Psm3 dimer, but Talon metal affinity resin (Clontech) was used for the histidine affinity purification.

Expression and purification of Psc3. The cDNA coding for Psc3, fused to a 6 × histidine tag at the N terminus, followed by a PreScission protease recognition sequence, was cloned into the pET30a bacterial expression vector. The resulting pET30a-Psc3 was transformed into *E. coli* BL21 (DE3) Rosetta. Fresh transformants were grown in Luria-Bertani medium containing 50 μg ml^{-1} kanamycin and 34 μg ml^{-1} chloramphenicol to OD_{600} = 0.5 at 37 °C. Isopropyl β-D-1-thiogalactopyranoside (IPTG; 0.5 mM) was added and cells were further grown at 18 °C for 18 h. Cells were collected and re-suspended in 10 pellet volumes of H buffer containing 300 mM KCl and protease inhibitors and disrupted by sonication. The lysate was clarified by centrifugation at 45,000*g* for 30 min at 4 °C and mixed with 1 ml Ni-NTA agarose slurry per 50 ml of the lysate for 4 h at 4 °C. The beads were washed with H buffer containing 300 mM KCl and bound proteins were eluted in H buffer containing 300 mM KCl and 250 mM imidazole. The eluate was adjusted to 100 mM KCl and loaded onto a HiTrap Heparin HP column that was developed with a linear gradient from 100 mM to 1 M KCl in H buffer. The peak fractions were pooled and applied to a Superdex 200 10/300 GL gel filtration column in H buffer containing 500 mM KCl. The peak fractions were dialysed against H buffer containing 200 mM KCl and concentrated to 500 μl by ultrafiltration. Typically, 0.1 mg of his-Psc3 was obtained from 2.5 l of culture, which was used for most of the experiments.

Alternatively, wild-type and mutant versions of Psc3 were expressed and purified using pGEX-6P-1 bacterial expression vector (GE Healthcare). Expression and lysate preparation were carried out as described above, except that 100 μg ml^{-1} ampicillin instead of kanamycin was used for cell culture. The lysate was incubated with glutathione sepharose 4B (GE Healthcare) at 4 °C for 4 h. Beads were then washed with H buffer containing 300 mM KCl and bound proteins were eluted with 20 mM glutathione in the same buffer. The eluate was treated with 5 U ml^{-1} PreScission protease at 4 °C overnight. The eluate was adjusted to 100 mM KCl in H buffer and loaded onto a HiTrap Heparin HP column. The column was developed with a linear gradient from 100 mM to 1 M NaCl in H buffer. The peak fractions were loaded onto a Superdex 200 10/300 GL gel filtration column that was developed in H buffer containing 200 mM KCl. The peak fractions containing Psc3 were concentrated to 500 μl. These preparations were used for the experiment shown in Fig. 5e and Extended Data Fig. 8b, c.

For use in the protein interaction studies shown in Fig. 5c and Extended Data Fig. 4c, we fused the *psc3*$^{+}$ cDNA C-terminally to a 3 × Pk followed by a 7 × histidine tag (Psc3-PkH) and cloned it into the expression vector pGEX-6P-1 to yield plasmid pGEX-Psc3PkH. Protein expression, cell lysate preparation and glutathione affinity purification were carried out as above. After PreScission protease treatment, the eluate was incubated with Ni-NTA agarose at 4 °C for 4 h. Beads were washed again with H buffer containing 300 mM KCl and the bound proteins were eluted in H buffer containing 300 mM KCl and 250 mM imidazole. The eluate was adsorbed to glutathione sepharose 4B to remove residual glutathione *S*-transferase, and the flow-through fraction was then dialysed against H buffer containing 300 mM KCl.

DNA, antibodies and peptides. The covalently closed circular (CCC) plasmid was pBluescript KSII (+). It was purified by equilibrium centrifugation in a CsCl-ethidium bromide gradient. rcDNA was prepared by treating the CCC form with *E. coli* topoisomerase I (New England BioLabs). Nicked circular DNA (NC) was prepared by DNase I treatment of the CCC preparation in the presence of ethidium bromide[41]. Linear DNA (L) was obtained from the CCC plasmid by restriction digestion using HindIII. Cohesin-bound DNA sequences from the fission yeast genome (E1, E2), and an equally sized control region (N; ref. 25), were amplified by PCR and cloned into the EcoRV restriction site in pBluescript KSII (+).

The coordinates of these regions on chromosome II were as follows: E1, 256411–261337; E2, 323792–328714; N, 244641–249740. The dsDNA and Y-fork structure DNA substrate for the electrophoretic mobility shift assays were prepared by annealing 50-nucleotide oligonucleotides, OL1+OL2 or OL1+OL3, respectively, followed by purification through polyacrylamide gel electrophoresis. The DNA sequences are listed in Supplementary Table 1. The sequences of the peptides used in the *in vitro* loading competition assay are found in Supplementary Table 2. The antibodies used for immunoprecipitation and western blotting were anti-Pk (V5, AbD Serotec), anti-HA (12C5A, Roche), anti-hexameric histidine (Novagen) and anti-tubulin (Cell Signaling). Antibodies raised against fission yeast Psm1, Rad21 and Mis4 were from BioAcademia.

DNA electrophoretic mobility shift assay. All indicated concentrations of the components are the final concentration in the reaction mixture. The 50-nucleotide DNA substrates were ^{32}P-labelled at their 5′ ends using polynucleotide kinase. 10 nM of the DNA substrates and the indicated concentrations of Mis4–Ssl3 were incubated in 25 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 50 mM NaCl, 0.1 mg ml^{-1} BSA and 8% (v/v) glycerol at 37 °C for 15 min. The reactions were then resolved using 5% acrylamide gel electrophoresis in 1 × TAE buffer at 4 °C. Gels were dried on DEAE paper. The gel image was visualized and band intensities were quantified using a Phosphorimager and ImageQuant software (GE Healthcare).

***In vitro* cohesin loading assay.** The standard reaction volume was 15 µl and contained 100 nM Mis4–Ssl3 and 3.3 nM molecules of rcDNA (pBluescript KSII (+); see above) which were mixed on ice in CL1 buffer (35 mM Tris/HCl, pH 7.5, 1 mM TCEP, 25 mM NaCl, 25 mM KCl, 1 mM $MgCl_2$, 15% (v/v) glycerol and 0.003% Tween 20). After 5 min, 150 nM cohesin and 100 nM Psc3 were added for further incubation on ice for 5 min. Free Psc3 was added to the reaction because Psc3 appeared somewhat sub-stoichiometric in our cohesin preparations and DNA binding by cohesin slightly increased by this addition (Fig. 2b and Extended Data Fig. 3d). The loading reaction was then initiated by addition of 0.5 mM ATP followed by incubation at 32 °C for 1 h. To stop the loading reaction and to dissociate non-topologically DNA-bound cohesin, 500 µl of CP buffer (35 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 500 mM NaCl, 10 mM EDTA, 5% (v/v) glycerol, 0.35% Triton X-100) was added to the reaction mixture and incubated at 32 °C for 5 min, followed by 5 min on ice. Anti-Pk-antibody-coated, protein-A-conjugated magnetic beads (Dynal Inc.) were added and rocked at 4 °C for 15 h. The magnetic beads were washed three times with CW1 buffer (35 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 750 mM NaCl, 10 mM EDTA, 0.35% Triton X-100), and then once with CW2 buffer (35 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 100 mM NaCl, 0.1% Triton X-100). The beads were then suspended in 15 µl elution buffer (10 mM Tris/HCl, pH 7.5, 1 mM EDTA, 50 mM NaCl, 0.75% SDS, 1 mg ml^{-1} protease K) and incubated at 50 °C for 20 min. The recovered DNA was analysed by 1% agarose gel electrophoresis in 1 × TAE and the gel was stained with GelRed (Biotium). Gel images were captured using a GelDoc-It Imager (UVP) and band intensities quantified using ImageQuant.

In experiments that included linearization of cohesin-bound cccDNA, the cohesin-bound DNA was retrieved by anti-Pk immunoprecipitation as described above, except that EDTA was omitted from the CP and CW buffers. The magnetic beads were further washed with RE buffer (35 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 100 mM NaCl, 10 mM $MgCl_2$, 0.1 mg ml^{-1} BSA, 0.1% Triton X-100). The beads were incubated with PstI (20 U, NEB) in 10 µl RE buffer at 10 °C for 3 h. The salt concentration was then adjusted to 500 mM NaCl in 15 µl and the reaction mixture was incubated on ice for 10 min, before the DNA molecules in the supernatant and bead fractions were analysed as described above.

The cleavage of engineered Rad21 by TEV protease was carried out at 16 °C for 2 h. TEV protease (10 U, Invitrogen) was added to the reaction mixture following the cohesin loading reaction. The cohesin-bound DNA was analysed as above.

The cohesin loading reactions using cohesin trimer and Psc3 mutants were carried out in the standard reaction condition with sight modification. Mis4–Ssl3 (100 nM) was initially mixed with rcDNA, and then 150 nM cohesin trimer and 150 nM Psc3 were added. The reaction mixture was incubated at the indicated temperature for 10 min, and then the reaction was initiated by addition of 0.5 mM ATP. After 1 h incubation at the indicated temperature, cohesin-bound DNA was analysed as described above.

For the peptide competition experiments, the cohesin loading reaction was carried out in CL2 buffer (50 mM Tris/HCl, pH 7.5, 1 mM TCEP, 25 mM NaCl, 1 mM $MgCl_2$, 15% (v/v) glycerol, 0.003% Tween 20). The indicated peptides (Psm1 peptides, 100 µM; Psm3 peptides, 100 µM; Psc3-1 peptides, 50 µM) were included with 50 nM Mis4–Ssl3 and 3.3 nM rcDNA for incubation on ice for 10 min, before 100 nM cohesin and 0.5 mM ATP were added to the reaction mixture for incubation at 32 °C for 45 min. Cohesin-bound DNA was then analysed as above.

ATPase assay. Cohesin (150 nM), 100 nM Mis4–Ssl3 and 3.3 nM rcDNA were mixed in CL1 buffer on ice. For the analyses of the cohesin trimer, 150 nM cohesin trimer, 100 nM Mis4–Ssl3, 150 nM Psc3 and 3.3 nM rcDNA were used. Note that no detectable ATP hydrolysis was observed in the absence of cohesin or cohesin trimer (Fig. 4c and data not shown). Reactions were initiated by addition of 0.25 mM ATP, spiked with [γ-^{32}P]-ATP (PerkinElmer), and incubated at 32 °C. Reaction aliquots were retrieved at 0, 15, 30 and 60 min and terminated by addition of three volumes of 500 mM EDTA. 1 µl of the terminated reactions were spotted on polyethylenimine cellulose F sheets (Merck), and separated by thin-layer chromatography using 400 mM LiCl in 1 M formic acid as the mobile phase. The separated spots representing ATP and released inorganic phosphate were quantified using a Phosphorimager and ImageQuant software.

Protein interaction analyses. For co-immunoprecipitation, 50 nM Mis4–Ssl3 and 50 nM cohesin or cohesin subunits (Psm1–Psm3 dimer, Psm3 or Psc3-PkH) were mixed in 50 µl of IP buffer (25 mM Tris/HCl, pH 7.5, 0.5 mM TCEP, 100 mM NaCl, 2.5 mM $MgCl_2$, 0.2% Triton X-100, 5% (v/v) glycerol) containing 0.5 mg ml^{-1} BSA and incubated at 25 °C for 15 min. After placing on ice for 15 min, the binding mixtures were transferred to anti-Pk-antibody-coated, protein-A-conjugated magnetic beads and rocked overnight at 4 °C. The beads were washed three times with IP buffer. The bound proteins were eluted in SDS–PAGE loading buffer, separated by 8.5% SDS–PAGE and detected by western blotting using anti-HA antibody to detect Mis4 and anti-Pk to detect Psm3 or Psc3. For far-western analysis, purified cohesin proteins were separated by SDS–PAGE, and transferred to a nitrocellulose membrane. The transferred proteins were re-natured in FW buffer (25 mM HEPES/KOH, pH 7.5, 150 mM KCl, 15 mg ml^{-1} BSA, 0.05% Tween 20) at room temperature (23 °C) for 2 h. The membranes were then blocked in FW buffer containing 5% milk powder. After rinsing with FW buffer, the membrane was incubated with 2.5 µg ml^{-1} Mis4–Ssl3 in FW buffer at 4 °C for 15 h. Bound Mis4–Ssl3 was detected by probing with an anti-HA antibody. For the interaction studies using tiling peptide arrays, 20-amino-acid-long peptides covering the amino acid sequences of Psm1, Psm3, Rad21 or Psc3, shifted by two amino acids, were synthesized on cellulose membranes using an Intavis Multipep peptide synthesizer (Intavis Bioanalytical Instruments AG). The membrane was activated in 50% methanol/10% acetic acid, and then blocked with 2.5% milk powder in TBS (25 mM Tris/HCl, pH 7.5, 150 mM NaCl, 0.1% Tween 20) at room temperature for 2 h. The blocked membrane was incubated with 1 µg ml^{-1} Mis4–Ssl3 in TBS buffer at 4 °C for 15 h. The membrane was washed with TBS and bound Mis4–Ssl3 was detected using an anti-HA antibody.

Fission yeast strain, media and genetic methods. Standard methods were used for fission yeast cell propagation and genetic manipulation, as described[42]. To investigate the consequence of mutations within the putative Mis4–Ssl3 interactions sites in cohesin, the *psm1*$^+$, *psm3*$^+$ and *psc3*$^+$ open reading frames were cloned behind their upstream promoter sequences (951 base pairs (bp), 977 bp and 990 bp, respectively) into the pARG1H vector[43]. At the C terminus, each protein was fused to a 3 × Pk followed by 7 × histidine tag to facilitate detection. The resultant plasmids formed the basis for the introduction of point mutations to alter Mis4–Ssl3 interaction sites using site-directed mutagenesis. Plasmids were linearized and integrated into the *arg1*$^+$ locus of the *psm1-897*, *psm3-602* and *psc3-303* temperature-sensitive strains, respectively, obtained from the Yeast National BioResource Project. Protein expression was monitored by western blotting of yeast whole-cell extracts obtained by glass bead breakage and TCA precipitation. For analysis of temperature sensitivity and drug resistance, exponentially growing cultures were spotted in fivefold serial dilutions on YES agar plates and grown at the indicated temperatures for 3 to 5 days.

Sister chromatid cohesion assay. The GFP-marked *cen2* locus[44] was introduced into the *psc3-303* strain, followed by the vectors carrying wild-type *psc3*$^+$ or the respective loader interaction site mutant alleles. Equal Psc3 expression levels were confirmed by western blotting. Exponentially growing cultures were shifted to 37 °C for 3 h. Cells were fixed with 70% ethanol. Cells were photographed and the percentage of cells with split GFP signals were counted using an Axioplan 2 Imaging microscope (Zeiss) equipped with a 100×/1.45 numerical aperture objective and an Orca-ER camera (Hamamatsu). At least 100 cells were scored for each strain, the experiment was performed three times, and the mean and standard deviations are shown.

40. Gietz, R. D. & Sugino, A. New yeast-*Escherichia coli* shuttle vectors constructed with *in vitro* mutagenized yeast genes lacking six-base pair restriction sites. *Gene* **74,** 527–534 (1988).
41. Shibata, T., Cunningham, R. P. & Radding, C. M. Homologous pairing in genetic recombination. Purification and characterization of *Escherichia coli* recA protein. *J. Biol. Chem.* **256,** 7557–7564 (1981).
42. Moreno, S., Klar, A. & Nurse, P. Molecular genetic analysis of fission yeast *Schizosaccharomyces pombe*. *Methods Enzymol.* **194,** 795–823 (1991).
43. Matsuyama, A., Shirai, A. & Yoshida, M. A novel series of vectors for chromosomal integration in fission yeast. *Biochem. Biophys. Res. Commun.* **374,** 315–319 (2008).

44. Yamamoto, A. & Hiraoka, Y. Monopolar spindle attachment of sister chromatids is ensured by two distinct mechanisms at the first meiotic division in fission yeast. *EMBO J.* **22,** 2284–2296 (2003).

45. Siegel, L. M. & Monty, K. J. Determination of molecular weights and frictional ratios of proteins in impure systems by use of gel filtration and density gradient centrifugation. Application to crude preparations of sulfite and hydroxylamine reductases. *Biochim. Biophys. Acta* **112,** 346–362 (1966).

ARTICLE

doi:10.1038/nature12928

UvrD facilitates DNA repair by pulling RNA polymerase backwards

Vitaly Epshtein[1]*, Venu Kamarthapu[1,2]*, Katelyn McGary[1], Vladimir Svetlov[1], Beatrix Ueberheide[1], Sergey Proshkin[3], Alexander Mironov[3,4] & Evgeny Nudler[1,2]

UvrD helicase is required for nucleotide excision repair, although its role in this process is not well defined. Here we show that *Escherichia coli* UvrD binds RNA polymerase during transcription elongation and, using its helicase/translocase activity, forces RNA polymerase to slide backward along DNA. By inducing backtracking, UvrD exposes DNA lesions shielded by blocked RNA polymerase, allowing nucleotide excision repair enzymes to gain access to sites of damage. Our results establish UvrD as a bona fide transcription elongation factor that contributes to genomic integrity by resolving conflicts between transcription and DNA repair complexes. Furthermore, we show that the elongation factor NusA cooperates with UvrD in coupling transcription to DNA repair by promoting backtracking and recruiting nucleotide excision repair enzymes to exposed lesions. Because backtracking is a shared feature of all cellular RNA polymerases, we propose that this mechanism enables RNA polymerases to function as global DNA damage scanners in bacteria and eukaryotes.

Nucleotide excision repair (NER) is the most versatile and evolutionarily conserved mechanism used by prokaryotic and eukaryotic cells to repair diverse types of DNA lesions[1,2]. In bacteria, the general NER pathway commences when UvrA and UvrB proteins bind damaged DNA and recruit UvrC to cleave the impaired strand on both sides of the lesion. The resulting oligonucleotide is displaced by UvrD and/or DNA polymerase I, which fills the gap using the complementary strand as a template[2–4].

NER rates are usually greatest at transcriptionally active genes. Moreover, the transcribed DNA strand is preferentially repaired compared to the non-transcribed strand[5]. This phenomenon, known as transcription-coupled repair (TCR), is a sub-pathway of global NER[3,6]. The current model of bacterial TCR postulates that a DNA lesion blocking the progression of the transcription elongation complex is shielded from NER enzymes by the stalled RNA polymerase (RNAP). The DNA translocase, Mfd, binds to the stalled elongation complex through the β subunit of RNAP and dislodges the complex by 'pushing' it forward[7–10]. Concurrently, Mfd recruits UvrA to the exposed lesion site to expedite NER[10].

Here we propose an alternative TCR model whose key component is UvrD, a member of DNA helicase superfamily 1, which translocates in a 3′ to 5′ direction using a single-strand, DNA-dependent, ATPase activity[11–13]. In contrast to Mfd, UvrD facilitates NER by pulling RNAP backward from the DNA lesion without causing termination. Our model further explains the role of elongation factor NusA, which is known to contribute to Mfd-independent TCR[14]. In this model RNAP recruits the NER complex via UvrD/NusA to the damage site.

UvrD binds RNAP *in vitro* and *in vivo*

We performed a mass spectrometry (MS)-assisted survey of proteins that interact with *E. coli* RNAP *in vivo* by treating *E. coli* K12 MG1655 cultures with formaldehyde and isolating RNAP-containing material. Peptides in this material were identified by tandem liquid chromatography mass spectrometry (LC-MS/MS) and used to calculate an exponentially modified protein abundance index (emPAI). This label-free method estimates the relative amount of proteins by the number of sequenced peptides per protein compared with the number of theoretically observable peptides[15]. UvrD appeared in RNAP crosslinked complexes in abundance comparable to that of bona fide transcription elongation/termination factors NusA, NusG or Rho (Extended Data Fig. 1), indicating a potential direct interaction between UvrD and RNAP.

To verify that UvrD, interacts with RNAP directly, we performed *in vitro* pull-down assays with purified UvrD and His6-tagged RNAP adsorbed to metal-chelating beads. UvrD bound to the beads only in the presence of immobilized RNAP and remained bound through multiple washings (Extended Data Fig. 1b). The UvrD–RNAP core complex was also isolated in a major discrete peak by size-exclusion chromatography (Extended Data Fig. 1c). Collectively, these data demonstrate stable and specific binding of UvrD to RNAP.

UvrD promotes RNAP backtracking

To investigate the role of UvrD in transcription we reconstituted a single-round runoff assay that measured 'walking' (nucleoside 5′-triphosphate (NTP) supply-controlled elongation) of the elongation complex along DNA[16]. We observed little effect of UvrD on RNAP promoter binding and open complex formation (not shown); however, UvrD profoundly influenced elongation, interrupting transcription at many positions along the template, so that only a fraction of elongation complexes produced a full-length (runoff) transcript (Fig. 1a, lane 2). Some UvrD-induced transcriptional 'arrests' coincided with pre-existing pause sites, whereas others formed *de novo*. Most did not change significantly with time, suggesting that the corresponding elongation complexes were either permanently arrested or terminated by UvrD. We excluded the latter possibility by showing that most of those transcripts remained bound to RNAP after extensive washing with high-salt buffer (Fig. 1a, lane 3).

Most transcriptional pauses and arrests are caused by backtracking—a reverse sliding of RNAP along DNA and RNA[17]. In bacteria, backtracked RNAP is prone to transcript cleavage stimulated by Gre factors,

[1]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York 10016, USA. [2]Howard Hughes Medical Institute, New York University School of Medicine, New York, New York 10016, USA. [3]State Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow 117545, Russia. [4]Engelhardt Institute of Molecular Biology, Russian Academy of Science, Moscow 119991, Russia.
*These authors contributed equally to this work.

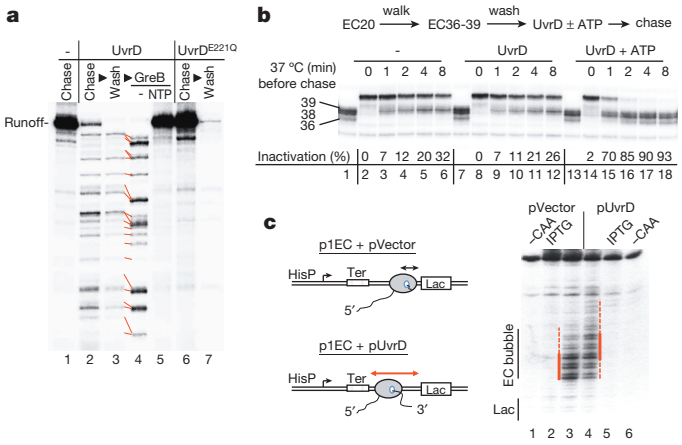


Figure 1 | UvrD promotes RNAP backtracking. **a**, EC20 was combined with wild-type UvrD (lanes 2–5), catalytically inactive UvrD^{E221Q} (lanes 6, 7), or no UvrD (lane 1) before NTP chase (lanes 1, 2, 6) and wash (lanes 3, 7). Red lines connect corresponding RNAs from arrested elongation complexes before and after GreB cleavage (lanes 4, 5). **b**, EC11 was walked to positions 36, 38 and 39. UvrD was added ± ATP and chased. Inactivated EC36–39 (%) is indicated. **c**, The p1EC constructs[21] (left) and primer extension analyses (right). CAA modifications on the non-template strand of p1EC + pVector (lanes 2, 3) or p1EC + pUvrD (lanes 4, 5). The lac operator (Lac) and transcription bubble are indicated. Red lines show elongation complex position.

which reactivate elongation complexes by removing the extruding 3′ portion of the nascent RNA[18,19]. To test whether elongation complexes arrested by UvrD were backtracked, we first washed them to remove unincorporated NTPs and UvrD, and incubated with GreB (Fig. 1a, lane 4). GreB shortened most transcripts from UvrD-arrested elongation complexes and reactivated these complexes in the presence of NTPs (lane 5), indicating that UvrD induces backtracking at many positions during elongation (Extended Data Fig. 2).

To determine whether the enzymatic activity of UvrD was required for the arrest, we repeated the above experiment with UvrD^{E221Q}; the E221Q mutation in the UvrD active site results in ~600-fold decrease in helicase/translocase activity without significantly affecting binding of DNA, ATP[20] or RNAP (not shown). UvrD^{E221Q} failed to cause RNAP arrest during elongation (Fig. 1a, lanes 6, 7).

To confirm the requirement of ATP for UvrD-mediated transcriptional arrest, we walked RNAP to stalled elongation positions (EC) 36–39 (counting from the +1 start of transcription), washed the beads to remove unincorporated NTPs and incubated the stalled complexes with UvrD ± ATP (Fig. 1b). After 1 min of incubation with UvrD + ATP most of EC36, EC38 and EC39 were inactivated: they failed to resume elongation upon addition of NTPs (lane 15); inactivation was almost complete after 2 min (lane 16). Without UvrD (lanes 2–6) or with UvrD lacking ATP (lanes 8–12), only ~30% of EC39 was inactivated after 8 min of incubation, whereas EC36 and EC38 remained fully active. We conclude that UvrD induces RNAP backtracking at many positions through its ATP-dependent motor function (Extended Data Fig. 2).

To monitor the effect of UvrD on RNAP backtracking *in vivo*, we used a plasmid (p1EC) in which RNA synthesis initiated at a constitutive promoter is halted at a downstream position by the *lacO*-bound Lac repressor (Fig. 1c). The plasmid was designed so that the repressor blocked one isolated elongation complex[21]. Cells carrying pEC1 were transformed with a UvrD overexpression plasmid (pUvrD) or an empty vector (pVector). To monitor the effect of UvrD and the position of the halted elongation complex we performed *in situ* footprinting of its DNA bubble using the single-strand-specific probe, chloroacetaldehyde (CAA) (Fig. 1c). The halted elongation complex was clearly backtracked over a longer distance in the presence of pUvrD than empty vector: new CAA reactive sites were detected upstream and the reactivity of the downstream margin of the footprint was decreased (lane 4). Thus, as observed *in vitro*, UvrD also causes RNAP to backtrack *in vivo*.

UvrD facilitates NER by towing RNAP

RNAP stalled at DNA lesions presents a major obstacle for NER by obscuring damaged sites from repair enzymes[1]. The role of UvrD, therefore, could be to clear such lesions by forcing the obstructing elongation complex to backtrack. To test this hypothesis we reconstituted the first steps of NER *in vitro* (Fig. 2a and Extended Data Fig. 3). An initial elongation complex was assembled on DNA carrying a single cyclobutane pyrimidine dimer (CPD) in the template strand at position +57 with respect to the transcriptional start site (Fig. 2a and Extended Data Fig. 3). CPD is the most common ultraviolet-induced lesion, a substrate for UvrABC and a roadblock for RNAP[22–24] (Extended Data Fig. 3). RNAP positioned 46 nucleotides upstream of CPD (EC11) did not affect UvrABC-directed CPD excision (Fig. 2a, lanes 3, 4, 7, 8). However, chasing EC11 to CPD inhibited the UvrC DNA cleavage reaction (lanes 11, 12). Addition of UvrD and ATP restored UvrC endonuclease activity at the CPD site (Fig. 2a, lanes 13–16), indicating that towing RNAP from the lesion site by UvrD is the first step of NER (Fig. 2a).

Anti-backtracking factors interfere with NER

In the RNAP 'towing' model of NER factors that normally inhibit backtracking may hinder the repair process: the GreA and GreB transcript cleavage factors[18] and active ribosomes control transcription elongation through this inhibition[21]. To test this prediction we used a primer extension assay to monitor ultraviolet-induced lesion repair *in vivo*. Such lesions block DNA polymerase, generating truncated primer extension products that can be analysed at single-nucleotide resolution[25]. We used a plasmid-borne *lacZ* gene isolated from wild-type and backtracking-prone *greA greB* mutant cells (Fig. 2b). The primer extension

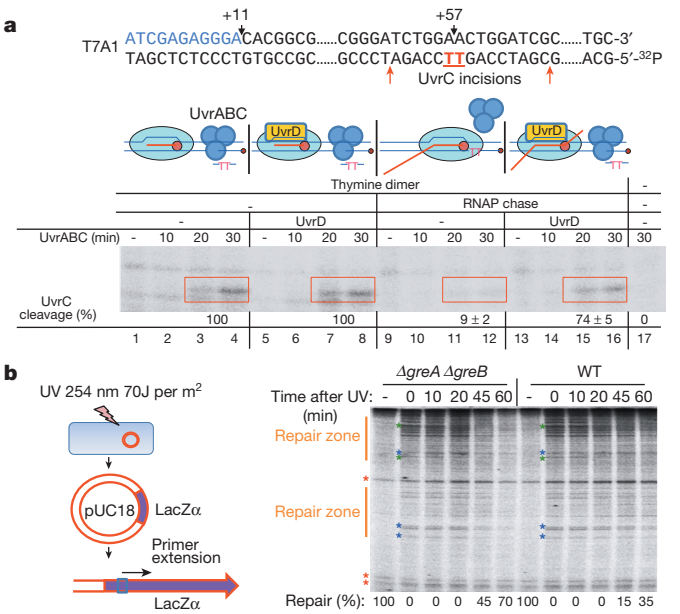


Figure 2 | RNAP backtracking facilitates NER. **a**, UvrD pulls RNAP from thymine dimers (TT). The T7A1 promoter template with TT (red) and schematic overview (top); EC11 was chased to TT (lanes 9–16) (bottom). Where indicated, UvrD was added for 5 min. Red boxes show UvrC-mediated DNA cleavage products (19 nucleotides); per cent cleavage is averaged (± s.e.m.) from four independent experiments $P < 0.05$. **b**, The primer extension experimental overview (left). As shown on the right, primer extension products reflect the location of lesions. Orange lines indicate repair zones for percentage repair. Asterisks indicate representative lesions repaired faster in *greAB* than in wild-type cells (blue), at the same rate (green) or 'non-disappearing' bands for normalization (orange). UV, ultraviolet; WT, wild type.

of *lacZ* DNA taken from both wild-type and mutant cells before and shortly after ultraviolet irradiation revealed truncated species unique to the irradiated DNA (Fig. 2b). However, the lesion repair rate (truncated species disappearance) was considerably higher in *gre*$^{-}$ cells than in wild-type cells. Only a few lesions were repaired at the same rate in wild-type and Gre-deficient cells. Because this assay detects any DNA damage that halts Taq DNA polymerase (not only pyrimidine dimers) we conclude that: (1) most ultraviolet-induced adducts within the transcribed region block the progression of RNAP, hampering the repair process, and (2) RNAP backtracking facilitates repair at most damage sites.

If these conclusions are correct, GreA and GreB deficiency should suppress UvrD sensitivity to ultraviolet and genotoxic agents against which NER provides protection. We examined three genotoxic chemicals: mitomycin C, 4-nitroquinoline-1-oxide (4NQO) and cisplatin. These chemotherapeutics generate crosslinks and bulky DNA adducts that are predominantly processed by NER[26]. *uvrD* cells were highly sensitive to killing by all three agents, whereas inactivation of *greAB* greatly suppressed *uvrD* sensitivity to all three agents (Fig. 3a and Extended Data Fig. 4) and to ultraviolet irradiation (Fig. 3b and Extended Data Fig. 4c). As the function of GreAB is to suppress RNAP backtracking, we conclude that RNAP backtracking is required for efficient NER *in vivo* and that UvrD is responsible for much of the NER-associated RNAP backtracking during genotoxic stress.

In bacteria, active ribosomes control the rate of transcription elongation at protein coding sequences by 'pushing' RNAP forward[21], whereas inhibited ribosomes render cells highly resistant to ultraviolet-mediated lethality[27,28]. To test if active ribosomes interfere with NER by decreasing the frequency of backtracking we used a sublethal dose of chloramphenicol to slow ribosomes' translocation[21]. Analogous to the situation with Gre-deficiency, chloramphenicol rendered *uvrD* cells more resistant to 4NQO, mitomycin (Fig. 3a and Extended Data Fig. 4b) and ultraviolet (Fig. 3b and Extended Data Fig. 4d).

Collectively, these results argue that UvrD competes with anti-backtracking factors during genotoxic stress to promote NER (Fig. 3c).

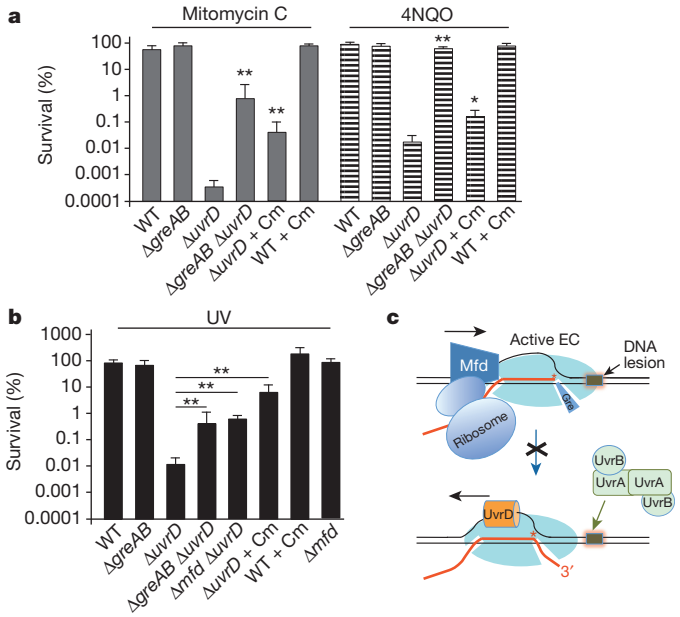


Figure 3 | Anti-backtracking factors obstruct UvrD activity in NER. **a**, Inactivating *greAB* or slowing ribosomal translocation (with 1 μg ml^{-1} chloramphenicol) suppresses *uvrD* sensitivity to mitomycin C (1 μg ml^{-1}; filled bars) and 4NQO (1 μM; striped bars). Data from three independent experiments are presented as the mean ± s.e.m.; $*P < 0.05$, $**P < 0.01$. Cm, chloramphenicol; WT, wild type. **b**, Inactivating *greAB*, *mfd* or slowing ribosomal translocation suppresses *uvrD* sensitivity to ultraviolet irradiation (5 J per m^2 at 30 °C). Data from three independent experiments are presented as the mean ± s.e.m.; $**P < 0.01$. **c**, Cartoon summarizing interference with NER.

Interestingly, UvrD counteracts GreAB even under normal growth conditions, because *uvrD* inactivation suppresses the temperature sensitivity of *gre*$^{-}$ cells (Extended Data Fig. 5).

Mapping UvrD–elongation complex interactions

We next mapped the location of UvrD within the elongation complex. Single-stranded/duplex DNA junctions are the preferred UvrD loading sites[29]; therefore, the upstream exposed portion of the transcription bubble[30] is an optimal UvrD binding site. In this model, UvrD bound to RNAP pulls the elongation complex backwards while unwinding the upstream fork of the bubble (Fig. 3c). To locate UvrD within the elongation complex we incorporated photo-inducible 4-thio-deoxyuridine-5′-monophosphate (4-thio-U) into the non-template strand at different positions (relative to catalytic site) to induce protein–DNA crosslinks. The −1 and −9 4-thio-U probes generated the most abundant crosslinked species corresponding to ββ′ subunits of RNAP; their pattern didn't change in the presence of UvrD (Fig. 4a, lanes 5–8). A unique UvrD-specific crosslinked species was detected only with the −9 probe (lane 6) and was consistent with the UvrD adduct. No significant crosslinks were detected in front of RNAP (position +22). These results support the model that UvrD binds near the upstream fork of the transcription bubble.

To further map interactions between UvrD and the elongation complex, we used bis[sulfosuccinimidyl] suberate (BS3) to crosslink proximal lysine residues between UvrD and RNAP. Using mass spectrometry we identified three inter-protein crosslinks between UvrD and RNAP. These crosslinks mapped to ββ′ subunits of RNAP and clustered around the DNA binding region on UvrD (Fig. 4b and Extended Data Fig. 6). In particular, one crosslink, (β K909)–(UvrD K124), mapped to the β flap tip domain. During elongation, the flap tip is required for activity of the NusA elongation factor[31]. We propose that UvrD also binds RNAP proximal to the flap tip domain in a way that allows UvrD to reach the non-template DNA strand near the upstream fork of the bubble. The positions of the remaining two crosslinks (β′ K79)–(UvrD K389) and (β′ K40)–(UvrD K448) are consistent with this hypothesis (Fig. 4b and Extended Data Fig. 6).

Considering the proximity of UvrD and NusA on the surface of RNAP, and the fact that NusA potentiates RNAP backtracking[32], we examined whether NusA supports UvrD-mediated backtracking. Indeed, NusA augmented UvrD-inducible arrests during elongation (Fig. 5a), providing a mechanistic explanation for the genetic evidence implicating NusA in Mfd-independent TCR[14]. Consistently, we showed that deletion of *greB* suppressed sensitivity of *nusA* cells (*ΔnusA* and *nusA11*) to 4NQO, mitomycin C, nitrofurazone (NFZ) and ultraviolet (Fig. 5b and Extended Data Fig. 7). It has been reported that UvrD and NusA directly bind UvrB and UvrA, respectively[14,33,34]. Thus, not only do the two elongation factors act synergistically to clear RNAP from the lesion site, they probably also recruit UvrAB to the damage site to facilitate repair (Fig. 5c).

RNAP as a global DNA damage surveillance vehicle

A major challenge for NER is that UvrAB must recognize almost every type of bulky adduct on DNA among millions of normal base pairs in the genome. The aberrations detected by UvrAB range from apurinic/apyrimidinic (AP) sites to ultraviolet-induced photoproducts to large chemical adducts[2]. The accuracy and efficiency of the repair of such variable lesion types would be improved by a screening mechanism that recognizes a common structure or process, such as that of arrested RNAP during transcription elongation.

The preferential repair of the template DNA strand, that is, TCR, implies that the transcription apparatus augments the search and/or binding of lesions by NER enzymes[3]. Cellular RNAPs are highly sensitive to aberrations in the template strand, and are temporally or permanently blocked by various DNA adducts[22–24,35]. As bacterial and eukaryotic genomes are pervasively transcribed, RNAP can serve as a global surveyor of DNA damage that constantly monitors the quality

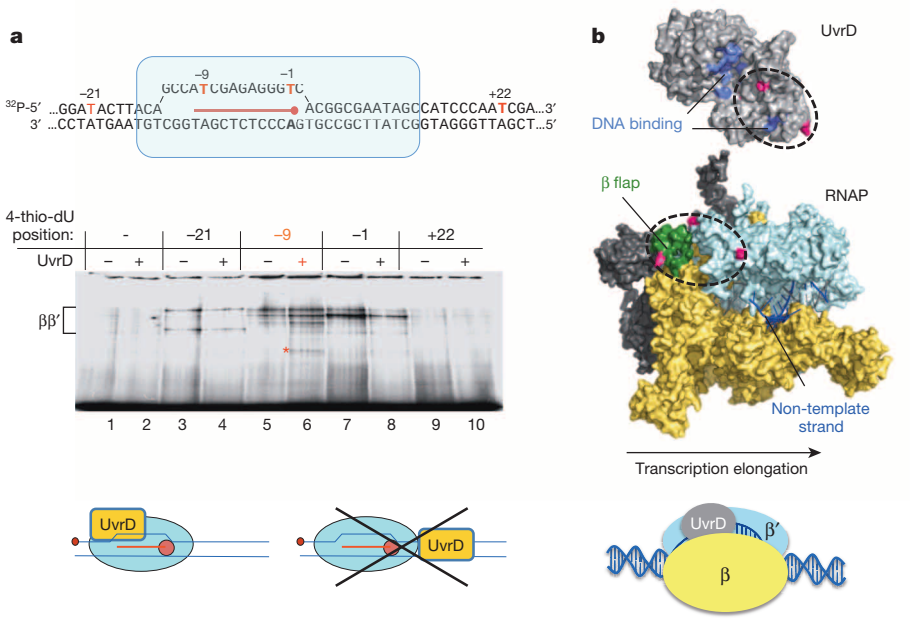


Figure 4 | Mapping UvrD interactions with the elongation complex. **a**, The 5′-radiolabelled scaffold carries a single photo-inducible 4-thio-dU (red) in the template strand (top). Protein crosslinking adducts corresponding to the β and β′ subunits of RNAP and UvrD (red asterisk) (middle). A model of UvrD binding (bottom). **b**, UvrD (top, Protein Data Bank (PDB) accession number 2IS4)[12] is crosslinked to RNAP (middle, PDB accession number 4IGC)[47] at three positions (magenta) that span the β (yellow) and β′ (light blue) subunits, proximal to the non-template strand (blue, PDB accession number 4G7O). The β flap-tip-helix (green) is indicated and the suggested binding interface is circled. Schematic summarizing UvrD–RNAP binding shown below.

of DNA via its natural one-dimensional diffusion. However, when stopped at DNA lesions, RNAP must either be terminated or moved aside to maintain damage site accessibility for NER. The Mfd pathway of TCR necessitates transcription termination[10], whereas the UvrD pathway described here uses RNAP backtracking. Importantly, this new UvrD mechanism facilitates NER without loss of RNAP, enabling transcription to promptly resume after DNA repair.

Mfd-independent TCR

The present work argues that the bulk of TCR occurs via UvrD-dependent backtracking (Fig. 5c); the sensitivity of *uvrD* and *nusA* cells to ultraviolet and various DNA damaging agents is much greater than that of *mfd* cells (Figs 3, 5b and Extended Data Figs 4, 7, 8). Yet, promoting backtracking by eliminating Gre factors, or by slowing ribosomal translocation, eliminates much of the DNA damage-induced

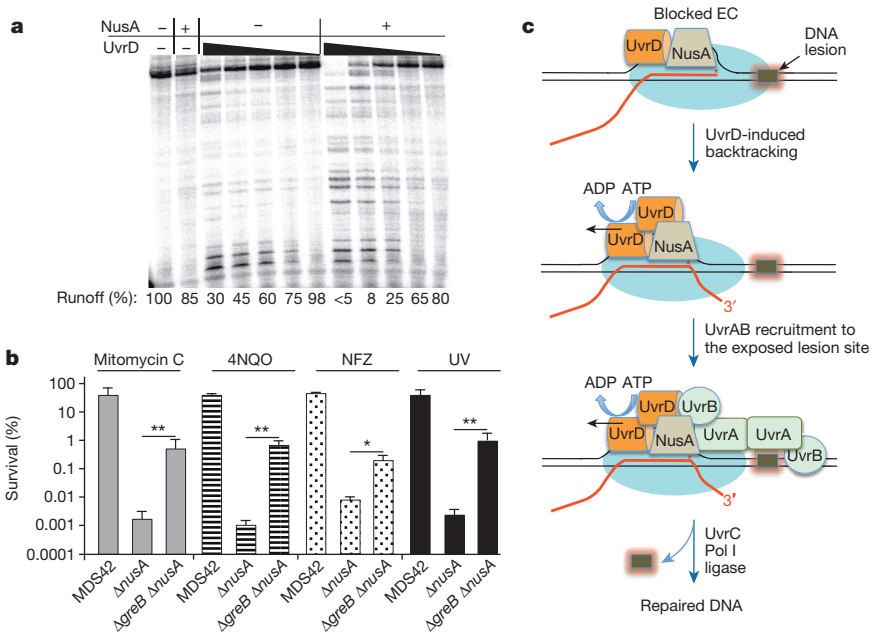


Figure 5 | UvrD and NusA cooperate in backtracking-mediated NER. **a**, NusA facilitates UvrD-mediated backtracking. The fraction of full-length transcript (percentage runoff) is indicated. **b**, GreB inactivation in MDS43 cells suppresses Δ*nusA* sensitivity to mitomycin C (0.5 μg ml^{-1}), 4NQO (3 μM), nitrofurazone (NFZ, 2 μM), and ultraviolet irradiation (15 J per m^2). Data from three independent experiments are presented as the mean ± s.e.m.; $*P < 0.05$, $**P < 0.01$. **c**, Model for backtracking-mediated NER.

lethality associated with UvrD or NusA deficiency (Figs 3, 5b and Extended Data Figs 4, 7). Consistently, deletion of DksA, a factor that competes with Gre for the secondary channel of RNAP, also increases *E. coli* sensitivity to mitomycin C[36]. Remarkably, deletion of *mfd* itself prominently suppresses *uvrD* sensitivity to ultraviolet and mitomycin C (Fig. 3b and Extended Data Figs 4, 8). This is consistent with the anti-backtracking mechanism of Mfd[8], further supporting the notion that RNAP backtracking is a prerequisite for most NER events (Fig. 3d). Survival after genotoxic challenge depends on both efficient NER and swift recovery from the SOS response. Indeed, *greAB* deletions facilitate NER, but do not increase cell survival (Fig. 3a, b). Similarly, Mfd function may not be as important for TCR per se, but instead for cellular recovery after the excessive backtracking associated with the SOS response.

In the absence of cellular stress, UvrD is equimolar with RNAP (~3,000 molecules per cell[37]). During the SOS response, however, the intracellular level of UvrD increases approximately threefold[11]. This spike facilitates UvrD dimerization and helicase activity[38] and probably is a prerequisite for UvrD-mediated backtracking. A sigmoid-shaped graph of RNAP arrest as a function of UvrD concentration (Extended Data Fig. 9) supports this view. After washing, UvrD remains bound to RNAP (Extended Data Fig. 1b), yet loses backtracking ability (Fig. 1a), suggesting monomeric and multimeric associations with RNAP during periods of no stress and stress, respectively (Fig. 5c). Excessive backtracking can be detrimental to genomic integrity in cells recovering from genotoxic stress and resuming replication, as frequent co-directional collisions between the replisome and backtracked elongation complexes result in dsDNA breaks (DSBs)[39]. By 'pushing' backtracked RNAPs forward, Mfd suppresses DSBs associated with such collisions[39], and hence diminishes the high frequency of mutations associated with DSBs repair. This model is consistent with the reduced 'mutation frequency decline' phenotype of *mfd* cells as well as their high ultraviolet mutability, minimal ultraviolet sensitivity[40], and compromised transcriptional recovery after ultraviolet exposure[41].

The process of TCR and the basic structural organization of RNAPs are evolutionarily preserved[1,6,26]. Bulky DNA lesions, such as thymine dimers, stall bacterial and eukaryotic RNAPs similarly[22,23,42]. As backtracking is a fundamental feature of all RNAPs[17] and occurs pervasively throughout the eukaryotic genome[43], it is likely to drive TCR in higher organisms as well. Notably, mammalian 3′–5′ DNA helicase, XPB, and its homologues from other eukaryotes, associates with elongating RNAP II as a subunit of the TFIIH transcription factor. It has been implicated in NER as well as in several human disorders associated with deficient DNA repair[44,45]. It is thus possible that the mechanistic role of XPB is analogous to that described here for UvrD, that is, backtrack-inducing TCR.

METHODS SUMMARY

Cultures of *E. coli* were crosslinked with formaldehyde, collected by centrifugation and lysed. RNAP-containing material was pulled down with anti-RNAP antibodies and proteins were identified by mass spectrometry. For DNA damage experiments, colony-forming units were counted after exposure to increasing concentrations of 4-NQO, mitomycin C or cisplatin. For ultraviolet survival assays, diluted cultures were plated, irradiated with a ultraviolet lamp and incubated overnight in the dark. Strains were transformed with pUC18 plasmid and induced with IPTG before and after irradiation. *In vitro* transcription assays were performed as described[16]. RNA–DNA scaffolds containing thymine dimers or 4-thio-dUMP-modified template oligonucleotides were assembled with immobilized, biotinylated RNAP before incubation with components of UvrABC system or irradiation at 308 nM for protein–DNA crosslinking, respectively. Purified RNAP and UvrD were crosslinked with BS3, and crosslinked proteins were reduced with dithiothreitol (DTT), alkylated with iodoacetamide and digested with trypsin. Tryptic peptides were analysed with a mass spectrometer coupled to a liquid chromatography system. pLink[46] was used to search Mascot generic format files for interpeptide crosslinks. For *in situ* DNA footprinting, the *uvrD* gene was cloned into a vector under the $P_{LtetO-1}$ promoter, and induced with anhydrotetracycline for 1 h before CAA modification. CAA modifications on non-template DNA were analysed with primer extension using [^{32}P]-labelled primer in parallel with the sequencing reactions.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 August; accepted 2 December 2013.
Published online 8 January 2014.

1. Reardon, J. T. & Sancar, A. Nucleotide excision repair. *Prog. Nucleic Acid Res. Mol. Biol.* **79,** 183–235 (2005).
2. Van Houten, B. & McCullough, A. Nucleotide excision repair in *E. coli*. *Ann. NY Acad. Sci.* **726,** 236–251 (1994).
3. Ganesan, A., Spivak, G. & Hanawalt, P. C. Transcription-coupled DNA repair in prokaryotes. *Prog. Mol. Biol. Transl. Sci.* **110,** 25–40 (2012).
4. Truglio, J. J., Croteau, D. L., Van Houten, B. & Kisker, C. Prokaryotic nucleotide excision repair: the UvrABC system. *Chem. Rev.* **106,** 233–252 (2006).
5. Mellon, I. & Hanawalt, P. C. Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. *Nature* **342,** 95–98 (1989).
6. Gaillard, H. & Aguilera, A. Transcription coupled repair at the interface between transcription elongation and mRNP biogenesis. *Biochim. Biophys. Acta* **1829,** 141–150 (2013).
7. Deaconescu, A. M. *et al.* Structural basis for bacterial transcription-coupled DNA repair. *Cell* **124,** 507–520 (2006).
8. Park, J. S., Marr, M. T. & Roberts, J. W. *E. coli* transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell* **109,** 757–767 (2002).
9. Savery, N. Prioritizing the repair of DNA damage that is encountered by RNA polymerase. *Transcription* **2,** 168–172 (2011).
10. Selby, C. P. & Sancar, A. Molecular mechanism of transcription–repair coupling. *Science* **260,** 53–58 (1993).
11. Kumura, K. & Sekiguchi, M. Identification of the *uvrD* gene product of *Escherichia coli* as DNA helicase II and its induction by DNA-damaging agents. *J. Biol. Chem.* **259,** 1560–1565 (1984).
12. Lee, J. Y. & Yang, W. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell* **127,** 1349–1360 (2006).
13. Matson, S. W. & George, J. W. DNA helicase II of *Escherichia coli*. Characterization of the single-stranded DNA-dependent NTPase and helicase activities. *J. Biol. Chem.* **262,** 2066–2076 (1987).
14. Cohen, S. E. *et al.* Roles for the transcription elongation factor NusA in both DNA repair and damage tolerance pathways in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **107,** 15517–15522 (2010).
15. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272 (2005).
16. Nudler, E., Gusarov, I. & Bar-Nahum, G. Methods of walking with the RNA polymerase. *Methods Enzymol.* **371,** 160–169 (2003).
17. Nudler, E. RNA polymerase backtracking in gene regulation and genome instability. *Cell* **149,** 1438–1445 (2012).
18. Borukhov, S., Sagitov, V. & Goldfarb, A. Transcript cleavage factors from *E. coli*. *Cell* **72,** 459–466 (1993).
19. Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89,** 33–41 (1997).
20. Brosh, R. M. Jr & Matson, S. W. Mutations in motif II of *Escherichia coli* DNA helicase II render the enzyme nonfunctional in both mismatch repair and excision repair with differential effects on the unwinding reaction. *J. Bacteriol.* **177,** 5612–5621 (1995).
21. Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328,** 504–508 (2010).
22. Donahue, B. A., Yin, S., Taylor, J. S., Reines, D. & Hanawalt, P. C. Transcript cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template. *Proc. Natl Acad. Sci. USA* **91,** 8502–8506 (1994).
23. Selby, C. P., Drapkin, R., Reinberg, D. & Sancar, A. RNA polymerase II stalled at a thymine dimer: footprint and effect on excision repair. *Nucleic Acids Res.* **25,** 787–793 (1997).
24. Selby, C. P. & Sancar, A. Transcription preferentially inhibits nucleotide excision repair of the template DNA strand *in vitro*. *J. Biol. Chem.* **265,** 21330–21336 (1990).
25. Manelyte, L., Kim, Y. I., Smith, A. J., Smith, R. M. & Savery, N. J. Regulation and rate enhancement during transcription-coupled DNA repair. *Mol. Cell* **40,** 714–724 (2010).
26. Batty, D. P. & Wood, R. D. Damage recognition in nucleotide excision repair of DNA. *Gene* **241,** 193–204 (2000).
27. Doudney, C. O. & Rinaldi, C. N. Chloramphenicol-promoted increase in resistance to UV damage in *Escherichia coli* B/r WP2 *trpE65*: development of the capacity for successful repair of otherwise mutagenic or lethal lesions in DNA. *Mutat. Res.* **143,** 29–34 (1985).
28. Hanawalt, P. C. The U.V. sensitivity of bacteria: its relation to the DNA replication cycle. *Photochem. Photobiol.* **5,** 1–12 (1966).
29. Tomko, E. J. *et al.* 5′-Single-stranded/duplex DNA junctions are loading sites for *E. coli* UvrD translocase. *EMBO J.* **29,** 3826–3839 (2010).

30. Korzheva, N. *et al.* A structural model of transcription elongation. *Science* **289,** 619–625 (2000).
31. Toulokhonov, I., Artsimovitch, I. & Landick, R. Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins. *Science* **292,** 730–733 (2001).
32. Bar-Nahum, G. *et al.* A ratchet mechanism of transcription elongation and its control. *Cell* **120,** 183–193 (2005).
33. Ahn, B. A physical interaction of UvrD with nucleotide excision repair protein UvrB. *Mol. Cells* **10,** 592–597 (2000).
34. Manelyte, L. *et al.* The unstructured C-terminal extension of UvrD interacts with UvrB, but is dispensable for nucleotide excision repair. *DNA Repair* **8,** 1300–1310 (2009).
35. Tornaletti, S. Transcription arrest at DNA damage sites. *Mutat. Res.* **577,** 131–145 (2005).
36. Trautinger, B. W., Jaktaji, R. P., Rusakova, E. & Lloyd, R. G. RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Mol. Cell* **19,** 247–258 (2005).
37. Arthur, H. M. & Eastlake, P. B. Transcriptional control of the *uvrD* gene of *Escherichia coli*. *Gene* **25,** 309–316 (1983).
38. Maluf, N. K., Fischer, C. J. & Lohman, T. M. A dimer of *Escherichia coli* UvrD is the active form of the helicase *in vitro*. *J. Mol. Biol.* **325,** 913–935 (2003).
39. Dutta, D., Shatalin, K., Epshtein, V., Gottesman, M. E. & Nudler, E. Linking RNA polymerase backtracking to genome instability in *E. coli*. *Cell* **146,** 533–543 (2011).
40. Witkin, E. M. Mutation and the repair of radiation damage in bacteria. *Radiat. Res.* **6,** (suppl.) 30–53 (1966).
41. Schalow, B. J., Courcelle, C. T. & Courcelle, J. Mfd is required for rapid recovery of transcription following UV-induced DNA damage but not oxidative DNA damage in *Escherichia coli*. *J. Bacteriol.* **194,** 2637–2645 (2012).
42. Brueckner, F. & Cramer, P. DNA photodamage recognition by RNA polymerase II. *FEBS Lett.* **581,** 2757–2760 (2007).
43. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).
44. Compe, E. & Egly, J. M. TFIIH: when transcription met DNA repair. *Nature Rev. Mol. Cell Biol.* **13,** 343–354 (2012).
45. Drapkin, R. *et al.* Dual role of TFIIH in DNA excision repair and in transcription by RNA polymerase II. *Nature* **368,** 769–772 (1994).
46. Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nature Methods* **9,** 904–906 (2012).
47. Murakami, K. S. X-ray crystal structure of *Escherichia coli* RNA polymerase σ70 holoenzyme. *J. Biol. Chem.* **288,** 9126–9134 (2013).

Acknowledgements We thank D. Jeruzalmi for materials. This work was supported by the Russian Foundation for Basic Research (A.M.) and the NIH, BGRF, Dynasty foundation and by the Howard Hughes Medical Institute (E.N.).

Author Contributions V.E., V.K., K.M., V.S., B.U., S.P. and A.M. conducted the experimental work, discussed the results and commented on the manuscript. E.N. designed the study and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.N. (evgeny.nudler@nyumc.org).

METHODS

Bacterial strains. The strains used in this study are listed in Extended Data Table 1 and were constructed using standard genetic techniques.

Mass spectrometry of *in vivo* RNAP interactome. Mid-exponential cultures of *E. coli* K12 MG1655 were treated with formaldehyde (Thermo Fisher, final concentration 50 mM) for 30 min, cells were collected by centrifugation, and lysed by a combination of ultrasound disruption and the action of lysozyme (human recombinant, Sigma-Aldrich). Pull-down of RNAP-containing material was performed using a mix of anti-RNAP antibodies (Neoclone) immobilized on Protein A/G Mag Sepharose beads (GE). Protein identification in the pull-down material was carried out at the NYUMC Protein Core Mass Spectrometry Facility, using an LTQ-Orbitrap mass spectrometer followed by Mascot database search (Matrix Science). Protein abundance was estimated using emPAI scores[15]. The range of emPAI values for a representative set of proteins in RNAP pull-downs spans 3 independent experiments (Extended Data Fig. 1).

Measuring sensitivity to DNA damage. *E. coli* strains were cultured in LB medium overnight. Appropriate dilutions were spread on LB agar plates containing increasing concentrations of 4NQO or NFZ or mitomycin C or cisplatin, followed by an overnight incubation at permissive temperatures. Colony-forming units (c.f.u.) were counted at each concentration of genotoxic agents. Chloramphenicol was used at the sublethal concentration (1.0 μg ml^{-1}), along with genotoxic agents. A 10 mM stock solution of NFZ or 4NQO was first made in *N,N*-dimethylformamide, stored at −20 °C, and diluted appropriately for each experiment. Mitomycin C was dissolved in sterile saline to a concentration of 500 μg ml^{-1} and stored at −20 °C. Cisplatin was dissolved in LB medium to a concentration of 500 μg ml^{-1} just before use. For ultraviolet survival assays, overnight grown cultures were diluted in M9 minimal salts, appropriate dilutions were spread on LB agar plates and irradiated with a ultraviolet (254 nm) lamp and incubated overnight in the dark at appropriate temperatures.

***In vivo* TCR.** Strains MG1655Z1 and MG1655Z1*ΔgreAΔgreB* were transformed with pUC18 plasmid. Overnight cultures were diluted 1:100 in LB medium supplemented with 100 μg ml^{-1} ampicillin and grown until $D_{600\,nm}$ ~0.3 at 30 °C. The culture was centrifuged at 5,000 r.p.m. for 5 min. The pellet was resuspended in M56 to an absorbance of 0.4.

The resuspended culture was spread on a Petri dish and, whilst being shaken, irradiated with 70 J per m^2 of 254 nm ultraviolet light. To induce the transcription on pUC18, 1 mM isopropyl-β-D-thiogalactoside (IPTG) was added to the culture medium 1 h before and immediately after ultraviolet irradiation. The irradiated culture was then supplemented with 0.25% casamino acids, 0.004% thiamine and 0.4% glucose and incubated at 30 °C in the dark to avoid removal of cyclobutane dimers by photoreactivation. Cells were collected before ultraviolet irradiation (-), immediately after (time 0 repair) and at defined repair times (15, 30, 45 and 60 min). Plasmid DNA was isolated from the cultures using a Qiagen plasmid isolation kit.

Primer extension was performed with [^{32}P]-labelled primers for analysis of cyclobutane pyrimidine dimers on the template strand. End labelled primer, 50 pmol, (5′-GGCATGCAAGCTTGGCACTGGC-3′), 400 μM dNTPs, 200 ng of plasmid DNA and Taq DNA polymerase were mixed in Thermopol buffer. The mixture was denatured by heating to 98 °C in a thermal cycler for 5 min followed by annealing at 55 °C for 5 min. Primer extension was performed at 68 °C for 6 min. Formamide stop solution was added to the reaction mixture and aliquots loaded onto a 6% urea-polyacrylamide gel. Percentages of repaired DNA were calculated using Image-Quant (GE) and averaged from three independent experiments with untreated samples taken as 100%.

***In vitro* transcription assays.** *E. coli* RNAP and UvrD were purified as described in refs 16 and 48, respectively. DNA templates were constructed using the non-transcribed part of T7 A1 promoter fragments fused to appropriate transcription units. Resulting DNA fragments were PCR amplified and purified from an agarose gel. For biotinylated DNA templates, biotinylated DNA oligonucleotides (IDT) were used. All NTP substrates were purified by ion-exchange chromatography[16]. Transcription at the T7A1-promoter templates was initiated by mixing 1 pmol of RNAP with equimolar amount of appropriate DNA PCR fragments in TB50 (TB, transcription buffer) (40 mM Tris-HCl pH 8.0; 10 mM $MgCl_2$; 50 mM NaCl) followed by addition of 10 μM ApUpC RNA primer, 25 μM ATP and GTP. Incubation was continued for 5 min at 37 °C. Resulting complexes were labelled by addition of 2 μCu α-[^{32}P]-CTP (3,000 Ci mmol^{-1}; MP Biomedicals) for 5 min at room temperature and immobilized on Neutravidin beads (5–10 μl; Piers) in the presence of 1.5 mg ml^{-1} heparin. The resulting EC20 was washed twice with 1 ml of TB1000 (TB with 1 M NaCl), twice with 1 ml of TB100 and divided into 10 μl aliquots. For the chase experiments, the EC20 was incubated with either the indicated amounts of UvrD or equal volumes of mock buffer for 5 min at room temperature and then chased for 5 min at 37 °C by addition of 0.1 mM NTPs. Where indicated, beads were washed after the chase with 1 ml of TB1000 and 1 ml of TB100. All reactions were stopped by equal volume of stop buffer (SB, 1×TBE buffer, 8 M urea, 20 mM EDTA, 0.025% xylene cyanol, 0.025% bromophenol blue). For the GreB cleavage assay (Fig. 2a), two 10 μl aliquots were incubated for 5 min at room temperature with 1 μM GreB. One aliquot was then chased for 5 min by addition of 0.1 mM NTPs and both samples were quenched as above. To test the effect of NusA on UvrD-induced backtracking (Fig. 5a), the initial EC20 was prepared as above and incubated with the indicated amounts of UvrD in the absence or presence of 10 nM NusA. Reaction mixtures were chased as above. For the UvrD-induced formation of stalled ECs (Fig. 2b), EC20 was walked to position 36–39 as described in ref. 16. The resulting ECs were mixed with 12.5 nM UvrD and 1 mM ATP or an equal amount of mock solution and incubated at 37 °C for the indicated time intervals. Next, 10 μl aliquots were withdrawn and mixed with 0.1 mM GTP and CTP for 5 min at room temperature. Reactions were stopped by addition of equal volumes of SB. The amount of UvrABC-mediated DNA cleavage products after 30 min at 37 °C was taken as 100% and UvrABC-mediated cleavage after CTP + GTP chase was calculated as a fraction of that value. Image-Quant (GE) software was used for all quantifications.

Reconstitution of the EC and UvrABC-mediated processing of thymine dimers. For assembly of the RNA–DNA scaffold (Fig. 2a), 75 pmol A1 RNA (IDT) were mixed with 75 pmol of the cyclothymine dimer template DNA strand or control strand (Midland) in 20 μl of the annealing buffer (AB) (12% glycerol; 20 mM Tris pH 8.0; 40 mM KCl; 5 mM $MgCl_2$) and annealed gradually in a PCR cycler. The resulting scaffold was stored at −20 °C and used as needed. To measure the extent of UvrABC-directed cleavage, 15 pmol of the DNA–RNA scaffold were mixed with 15 pmol of biotinylated RNAP in 20 μl of AB and incubated for 10 min at room temperature. 75 pmol of non-template strand was added and incubation continued for another 5 min. The resulting complex was immobilized at Neutravidin-coated beads (Pierce), washed twice with TB1000 and then with TB100. For RNA labelling, 5 μM ATP and 2 μCu α-[^{32}P]-CTP were added for 5 min at room temperature, followed by washing with TB100. For the UvrC cleavage reaction complexes were washed twice with TB0 (20 mM Tris-HCl pH 8.0; 10 mM $MgCl_2$) and 10× PNK buffer (New England Biolabs) was added together with 1 μCu γ-[^{32}P]-ATP (7,000 Ci mmol^{-1}; MP Biomedicals) and 20 units PNK (New England Biolabs). The DNA labelling reaction continued for 30 min at 37 °C, then the beads (EC11) were washed twice with TB1000 and TB100. EC11 were processed directly or chased first for 10 min at 37 °C by the addition of 1 mM NTPs. Complexes were supplemented with 1 mM ATP and 12.5 nM UvrD was added where indicated for 10 min at 37 °C before washing with TB100. To initiate the cleavage reaction premixed UvrA (2.5 nM), UvrB (10 nM) and UvrC (25 nM) were added and incubated for the indicated time intervals at 37 °C before quenching with SB. UvrABC were purified as described in ref. 49. Products of all reactions were separated in 6–23% denaturing polyacrylamide gels and visualized via phosphor-imaging using a Typhoon phosphorimager (GE Healthcare) and Image Quant software (GE Healthcare).

Protein–DNA crosslinking. RNA–DNA scaffolds were assembled as above using the 4-thio-dUMP-modified template oligos (Midland), A1 RNA (9 nt long) and non-template DNA. Resulting complexes were labelled as described above and incubated with 12.5 nM UvrD or mock solution for 5 min at 37 °C before being irradiated for 10 min with a hand-held ultraviolet lamp (Cole Parmer) at 308 nM wavelength. Irradiation was performed on ice and samples were denatured in Laemmli loading buffer. Products were separated in a 6% denaturing polyacrylamide gel with 0.1% SDS and Tris-glycine running buffer. UvrD adducts were seen on the gel as a 90 kDa species.

UvrD–RNAP binding assay. UvrD and RNAP, UvrD alone, and RNAP alone were incubated at room temperature for 30 min in binding buffer (50 mM Tris pH 8.0, 100 mM NaCl, 50 mM imidazole) before they were added to His Mag Sepharose Ni beads (GE), which were pre-equilibrated and washed with binding buffer. Each was then incubated for 30 min with gentle agitation at room temperature. After incubation, beads were washed three times with binding buffer and eluted in 50 μl of elution buffer (50 mM Tris pH 8.0, 100 mM NaCl, 400 mM imidazole). Samples were electrophoresed in an SDS–polyacrylamide gel alongside pure UvrD, for reference.

RNAP–UvrD crosslinking and their mapping by LC-MS/MS. Purified RNAP at a final concentration of 2 μM was incubated with 4 μM UvrD in crosslinking buffer (1× PBS, 500 mM NaCl, pH 7.4) for 30 min at room temperature. Crosslinking was initiated upon addition of Bis(sulfosuccinimidyl) suberate (BS3) (Thermo Scientific) in 1×PBS at a final concentration of 350 μM. Crosslinking reactions continued for 30 min at room temperature before they were quenched with 1 M Tris pH 8.0 at a final concentration of 200 mM.

Crosslinked proteins were separated by SDS–PAGE (4–12% Bis Tris Novex gels, Invitrogen) and stained with GelCode Blue. Super-shifted (relative to positions of untreated RNAP subunits) were removed from the gel and destained. Samples were reduced with 20 mM DTT (Sigma) at 57 °C for 60 min and alkylated with

45 mM iodoacetamide (Sigma) in the dark for 45 min at room temperature before overnight, in-gel digestion with 0.5 μg sequencing grade modified trypsin (Promega).

Gel slices were incubated with Poros R2 50 μm slurry diluted in 5% formic acid and 0.2% TFA to bind tryptic peptides. The Poros bead slurry was collected and washed through C18 Zip Tips (Millipore) before elution with 40 μl of 40% acetonitrile in 0.5% acetic acid. Eluted peptides were dehydrated in vacuum and resuspended in 40 μl 0.5% acetic acid for MS analysis.

Peptide aliquots (~200 ng) were analysed in the Q Exactive mass spectrometer (Thermo Scientific) coupled to an EASY-nLC (Thermo Scientific) liquid chromatography system. The peptides were eluted over a 120-min linear gradient from 98% buffer A (water) to 40% buffer B (ACN) then continue to 100% buffer B over 10 min with a flow range of 200 nl min^{-1}. Each full MS scan (R = 70,000) was followed by dd-MS^2 (R = 17,500) with HCD and an isolation window of 2.0 *m/z*. Normalized collision energy was set to 35. Precursors of +1, +2 and +3 were excluded from MS2 scans; monoisotopic screening was enabled and a dynamic exclusion window was set to 30.0 s.

For identification of crosslinks, mascot generic format (mgf) files were generated from Xcalibur raw files using the Proteome Discoverer software (Thermo Scientific). pLink[46] was used to search resulting mgf files for interpeptide crosslinks with the following settings: precusor mass tolerance 20 p.p.m.; fragment mass tolerance 5 p.p.m.; BS3 crosslinker (crosslink monoisotopic mass shift of 138.0680786 Da, monolink mass shift of 156.0786442 Da). One fixed modification (Carbamidomethyl-C) and three variable modifications (oxidation-M, N-terminal glutamate to pyroglutamate, N-terminal acetylation) were specified. False discovery rate was set at <5%. Fragmentation spectra for each crosslinked peptide were validated using pLabel[50] and confirmed with manual inspection.

***In situ* DNA footprinting.** To construct the plasmid pUvrD (Cm^r, p15a), the *uvrD* gene was excised from the plasmid pETDuet-uvrD as an XhoI / filled-in XbaI fragment and cloned into the pZA31 vector at the SalI site (compatible cohesive end with XhoI) and the blunted KpnI site. The expression of the *uvrD* gene in the resulting plasmid is driven by the P_L-derivative promoter $P_{LtetO-1}$ which is controlled by the operator repressor system of the Tn*10*-derived *tet* resistance operon[51]. The induction of $P_{LtetO-1}$ is achieved by anhydrotetracycline.

To ensure expression of the regulatory repressor proteins LacI and TetR, the entire unit encoding LacR, TetR and Sp^r[53] was transferred into the chromosome of *E. coli* strain MG1655 by P1 transduction (MG1655Z1).

E. coli strains were grown in 10 ml of M9 medium supplemented with 0.4% glucose, 4 mg ml^{-1} casamino acids, 100 μg ml^{-1} ampicillin and 30 μg ml^{-1} chloramphenicol. To induce the $P_{LtetO-1}$ promoter, anhydrotetracycline (50 ng ml^{-1}) was added for 1 h before CAA modification. At $D_{600\,nm}$ ~ 0.6, 1 mM IPTG was added (if required) for 10 min, then CAA was added to 4% final concentration and the incubation was continued for 7 min. Then cells were collected, immediately washed with the saline solution, collected and the plasmid DNA extracted.

To analyse CAA modification[52] of the non-template DNA strand, primer extension was performed with [^{32}P]-labelled primer (5′-TAGCTTCCTTAGCT CCTGA-3′) in parallel with the sequencing reactions in a cycler at the following conditions: 94 °C for 3 min (initial denaturation), then 94 °C for 30 s (denaturation), 56 °C for 30 s (annealing), 72 °C for 60 s (extension), 25 cycles total. Formamide stop solution was added to the reaction mixture and aliquots were analysed by electrophoresis in an 8% urea-polyacrylamide gel.

48. Runyon, G. T., Wong, I. & Lohman, T. M. Overexpression, purification, DNA binding, and dimerization of the *Escherichia coli uvrD* gene product (helicase II). *Biochemistry* **32,** 602–612 (1993).
49. Pakotiprapha, D., Samuels, M., Shen, K., Hu, J. H. & Jeruzalmi, D. Structure and mechanism of the UvrA-UvrB DNA damage sensor. *Nature Struct. Mol. Biol.* **19,** 291–298 (2012).
50. Li, D. *et al.* pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **21,** 3049–3050 (2005).
51. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25,** 1203–1210 (1997).
52. Epshtein, V., Toulme, F., Rahmouni, A. R., Borukhov, S. & Nudler, E. Transcription through the roadblocks: the role of RNA polymerase cooperation. *EMBO J.* **22,** 4719–4727 (2003).
53. Pósfai, G. *et al.* Emergent properties of reduced-genome *Escherichia coli. Science* **312,** 1044–1046 (2006).
54. Cardinale, C. J. *et al.* Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in *E. coli. Science* **320,** 935–938 (2008).

LETTER

doi:10.1038/nature12916

A Be-type star with a black-hole companion

J. Casares[1,2], I. Negueruela[3], M. Ribó[4], I. Ribas[5], J. M. Paredes[4], A. Herrero[1,2] & S. Simón-Díaz[1,2]

Stellar-mass black holes have all been discovered through X-ray emission, which arises from the accretion of gas from their binary companions (this gas is either stripped from low-mass stars or supplied as winds from massive ones). Binary evolution models also predict the existence of black holes accreting from the equatorial envelope of rapidly spinning Be-type stars[1–3] (stars of the Be type are hot blue irregular variables showing characteristic spectral emission lines of hydrogen). Of the approximately 80 Be X-ray binaries known in the Galaxy, however, only pulsating neutron stars have been found as companions[2–4]. A black hole was formally allowed as a solution for the companion to the Be star MWC 656 (ref. 5; also known as HD 215227), although that conclusion was based on a single radial velocity curve of the Be star, a mistaken spectral classification[6] and rough estimates of the inclination angle. Here we report observations of an accretion disk line mirroring the orbit of MWC 656. This, together with an improved radial velocity curve of the Be star through fitting sharp Fe II profiles from the equatorial disk, and a refined Be classification (to that of a B1.5–B2 III star), indicates that a black hole of 3.8 to 6.9 solar masses orbits MWC 656, the candidate counterpart of the γ-ray source AGL J2241+4454 (refs 5, 6). The black hole is X-ray quiescent and fed by a radiatively inefficient accretion flow giving a luminosity less than 1.6×10^{-7} times the Eddington luminosity. This implies that Be binaries with black-hole companions are difficult to detect in conventional X-ray surveys.

The majority of Be X-ray binaries[7] (BeXBs) contain proven neutron stars and are characterized by their transient changes in X-ray luminosity, when episodes of increased accretion onto the compact star are modulated either by the periastron passage or tidal disruption of the Be circumstellar disk[8]. When in quiescence, they have very low (or even undetectable) X-ray emission. It has been proposed that BeXBs with black holes (Be–BH) are difficult to find because of efficient disk truncation, leading to very long quiescent states[2]. Alternatively, their absence could be driven by binary evolution, with Be–BH binaries having a lower probability of being formed and surviving a common envelope phase[3].

MWC 656 is a Be star located within the error box of the point-like γ-ray source AGL J2241+4454 (ref. 9). A photometric modulation of 60.37 ± 0.04 days was reported, suggesting that MWC 656 is a member of a binary[6]; this was subsequently confirmed through radial velocities of He I lines from the photosphere of the Be star[5]. The radial velocity curve, though, displays considerable scatter, likely to be caused by filled-in emission from the circumstellar wind contaminating the broad absorption profiles (a common limitation in the analysis of BeXBs; see, for example, ref. 10) and only a tentative orbital solution is available[5]. Here we revisit the 32 Liverpool telescope spectra previously reported[5]. These are complemented with 4 additional Liverpool telescope spectra and a further high-resolution echelle spectrum obtained with the 1.2-m Mercator telescope (see Methods and Extended Data Table 1).

A close-up of the Mercator telescope spectrum is presented in Fig. 1, showing classic Fe II emission lines from the Be circumstellar disk. In addition, a He II 4,686 Å emission line (which was overlooked in a previous work[5]) stands out clearly: its presence is remarkable because it requires temperatures hotter than can be achieved in disks around B-type stars. Further, the He II profile is double-peaked, which is the signature of gas orbiting in a Keplerian geometry[11]. Gaussian fits to the He II profiles in the Liverpool telescope spectra reveal that the centroid of the line is modulated with the 60.37-day orbital period, reaching maximum velocity at photometric phase 0.06 (see Methods and Extended Data Fig. 1). This is approximately in antiphase with the radial velocity curve of the Be star[5], a strong indication that the He II emission arises from gas in an accretion disk around the invisible companion and not from the Be disk. We can therefore use its radial velocity curve to trace the orbit of the Be companion. An eccentric orbital fit to the He II velocities was performed using the Spectroscopic Binary Orbit Program (SBOP[12]), fixing the period to 60.37 days (Methods); the resulting orbital elements are given in Extended Data Table 2. The orbital evolution of the He II line is presented in Fig. 2. The line flux is also found to be modulated with the orbital period (Methods and Extended Data Fig. 1), owing to the presence of an S-wave component swinging between the double peak (see Fig. 2).

To improve on the radial velocity curve of the Be star previously reported[5], we fitted the sharp double-peaked profile of the Fe II 4,583 Å emission line with a two-Gaussian model (Methods). Fe II lines are known to arise from the innermost regions of the circumstellar disk[13,14], and therefore reflect the motion of the Be star much more accurately

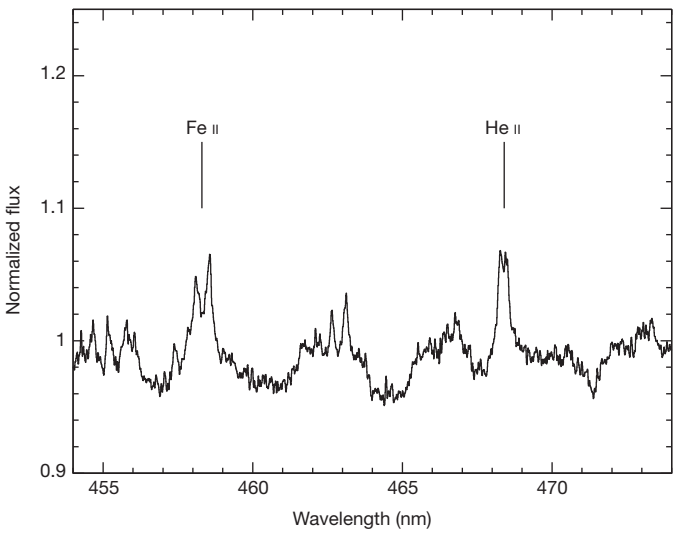


Figure 1 | Optical spectrum of MWC 656, obtained with the Mercator telescope. The spectrum has been rebinned to 0.1 Å per pixel and smoothed through a 3-pixel Gaussian bandpass. The emission lines Fe II 4,583 Å (of multiplet 38) and He II 4,686 Å are indicated. Whereas Fe II is formed in the Be circumstellar disk, He II arises from gas encircling the companion star. Several other circumstellar Fe II lines are detected in the spectrum (for example, 4,549, 4,555, 4,629, 4,666, 4,731 Å) but they are weaker and/or severely blended with other emission lines and photospheric absorptions.

[1]Instituto de Astrofísica de Canarias, E-38205 La Laguna, Santa Cruz de Tenerife, Spain. [2]Departamento de Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Santa Cruz de Tenerife, Spain. [3]Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal, Universidad de Alicante, Apartado 99, E-03080 Alicante, Spain. [4]Departament d'Astronomia i Meteorologia, Institut de Ciències del Cosmos, Universitat de Barcelona, IEEC-UB, Martí i Franquès 1, E-08028 Barcelona, Spain. [5]Institut de Ciències de l'Espai—(IEEC-CSIC), Campus UAB, Facultat de Ciències, Torre C5 - parell - 2a planta, E-08193 Bellaterra, Spain.

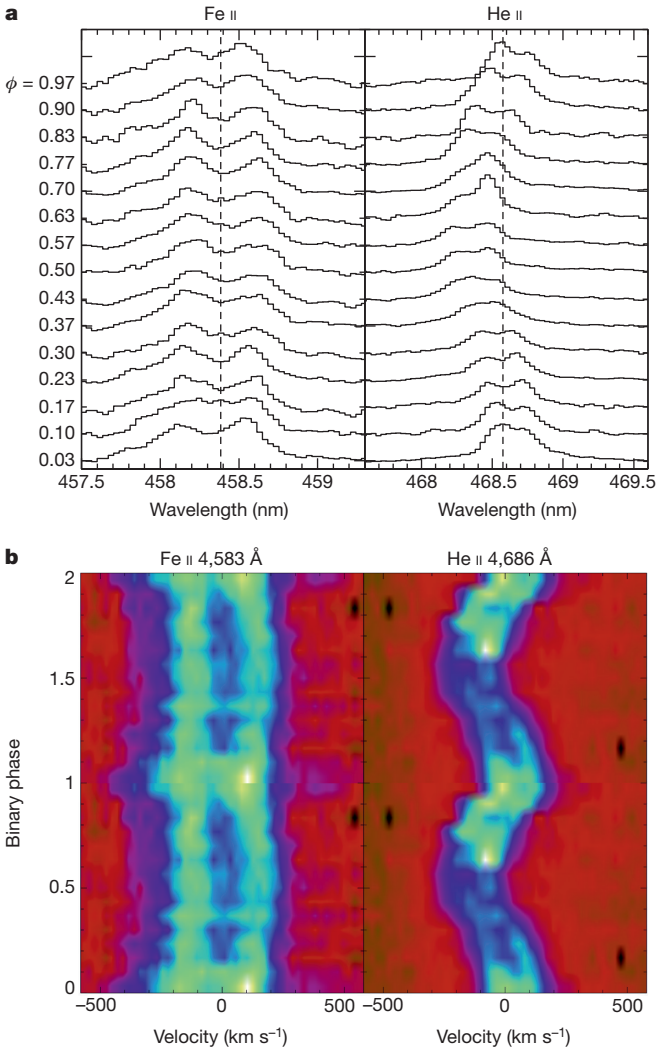


Figure 2 | Orbital evolution of the Fe II 4,583 Å and He II 4,686 Å emission lines. **a**, Sequence of spectra folded into 15 phase bins using the photometric ephemeris[6]. ϕ indicates the binary orbital phase. Vertical dashed lines indicate the central wavelength of each line. For the sake of clarity, the flux of the Fe II line has been arbitrarily scaled by a factor of 2 with respect to He II. The He II double peak is distorted by an S-wave component, typically associated with a bright spot or asymmetry in the outer accretion disk[22]. **b**, Trailed intensity image of the two lines constructed from the phase binned spectra. Two orbital cycles are displayed for clarity. The colour scale indicates counts normalized to the continuum, with the black colour corresponding to 0.98 and the white colour to 1.08 in Fe II and 1.16 in He II.

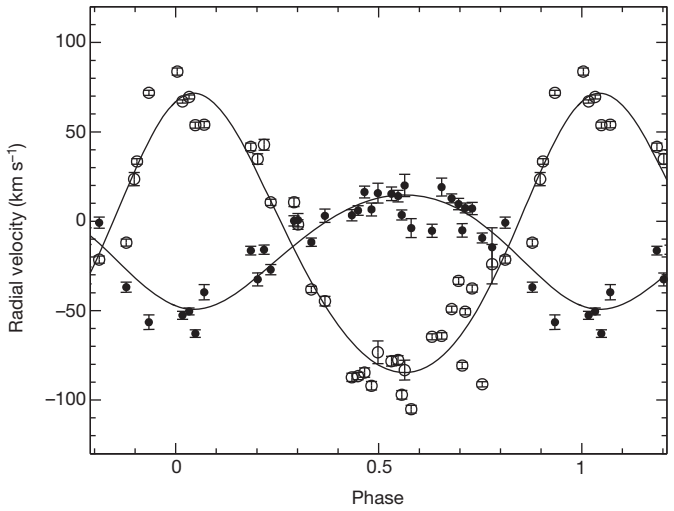


Figure 3 | The radial velocity curves of the Be star and its companion. Filled circles indicate radial velocities of the Be star, as obtained from the Fe II 4,583 Å line; open circles indicate those of the companion star, extracted from the He II 4,686 Å line. Error bars, 1 s.d. The best combined orbital fit is overplotted. The velocities have been folded with the photometric ephemeris[6] as in Fig. 2. A shift of +30 km s^{-1} was applied to the He II velocities to compensate for the S-wave contamination (Methods and Extended Data Fig. 2). Two Fe II velocities (not shown) were found to deviate by $>2\sigma_f$ (where σ_f is the r.m.s. of the fit) from the best solution and were excluded from the fit, but they have a negligible impact on the final orbital elements.

than the broad He I absorptions. The Fe II velocities were also modelled with SBOP, fixing the period to 60.37 days. The eccentric orbital fit results in orbital elements that are consistent with the Fe II orbit being the reflex of the He II orbit, as expected from the motion of two components in a binary system (see Methods and Extended Data Table 2). Only the eccentricities are slightly discrepant, but just at the 1.6σ level.

Consequently, we modelled the ensemble of Fe II and He II radial velocities with a double-line eccentric binary orbit in SBOP. Figure 3 presents the radial velocity curves of the two emission lines with the best combined solution superimposed. The resulting orbital elements are listed in Table 1. Our solution yields a mass ratio $q = M_2/M_1 = 0.41 \pm 0.07$, which implies a rather massive companion star. A precise determination of the companion's mass requires an accurate spectral classification of the Be star. Accordingly, we have compared the Mercator telescope spectrum with a collection of observed B-type templates, broadened by 330 km s^{-1} to mimic the large rotation velocity in MWC 656 (Methods and Extended Data Fig. 3). Using excitation temperature diagnostics based on several absorption line ratios, we determine a spectral type of B1.5–B2, and the strength of the metallic lines combined with the moderate width of the Balmer absorption wings implies luminosity class III (Methods). Our adopted B1.5–B2 III classification implies a mass of 10–16 solar masses ($M_\odot$) for the Be star (Methods), and hence a companion star of 3.8–6.9 $M_\odot$.

The large dynamical mass of the companion to the Be star MWC 656 is puzzling. A normal main-sequence star with such mass would have a spectral type in the range B3–B9 and its spectrum would be easily detected in the optical range. Nor can it be a subdwarf, because these typically have masses in the range 0.8–1.3 $M_\odot$ (ref. 15). The stripped He core of a massive progenitor (that is, a Wolf–Rayet star) is also rejected because it possesses strong winds which show up through intense high-excitation emission lines, not present in our spectra. In

Table 1 | Orbital elements for MWC 656

Parameter	Value
P_{orb} (days)	60.37 (fixed)
T_0 (HJD − 2,450,000)	3,243.70 ± 4.30
e	0.10 ± 0.04
ω (degrees)	163.0 ± 25.6
γ (km s^{-1})	−14.1 ± 2.1
K_1 (km s^{-1})	32.0 ± 5.3
K_2 (km s^{-1})	78.1 ± 3.2
$a_1 \sin i$ ($R_\odot$)	38.0 ± 6.3
$a_2 \sin i$ ($R_\odot$)	92.8 ± 3.8
$M_1 \sin^3 i$ ($M_\odot$)	5.83 ± 0.70
$M_2 \sin^3 i$ ($M_\odot$)	2.39 ± 0.48
M_2/M_1	0.41 ± 0.07
σ_f (km s^{-1})	16.7

The solution was obtained from a combined fit to the radial velocity curves of the He II 4,686 Å and Fe II 4,583 Å lines. The orbital period P_{orb} has been fixed to the photometric value[6]. T_0 is the epoch of periastron (where HJD refers to heliocentric Julian date), e the orbit eccentricity, ω the longitude of periastron, γ the systemic velocity, K the velocity semiamplitude, a the semimajor axis, i the binary inclination, M the stellar mass and σ_f the r.m.s. of the fit. Subscripts 1 and 2 refer to the primary (Be star) and secondary (companion) components, respectively. T_0 implies that periastron passage occurs at photometric phase 0.01 ± 0.10.

addition, this should be detected as an ultraviolet excess in the spectral energy distribution, which is not observed in the fluxes available in the literature (Methods). On the other hand, the evidence of a He II accretion disk encircling the companion star strongly points towards the presence of a compact object. The large dynamical mass rules out a white dwarf or a neutron star, so the only viable alternative is a black hole. It should be noted that none of the ~170 BeXBs currently known[4] shows any evidence for an accretion disk, providing circumstantial evidence for a difference in the nature of the compact stars. The accretion disk in MWC 656 is expected to also radiate Balmer and He I lines but these are blended with the corresponding (stronger) emission lines from the Be disk and thus are not detected.

MWC 656 is a key system in the study of BeXBs and massive binary evolution. At a distance $d = 2.6 \pm 0.6$ kpc (Methods) it is relatively nearby and also one of the visually brightest Be binaries[7]. It thus seems reasonable to assume that many other Be–BHs exist in the Galaxy but remain hidden by the lack of transient X-ray activity. Analysis of archival ROSAT images yields an upper limit to the X-ray flux at energies of 0.1–2.4 keV of 1.2×10^{-13} erg cm^{-2} s^{-1} (Methods) which, for our estimated distance, translates into an X-ray luminosity $L_X < 1.0 \times 10^{32}$ erg s^{-1} or $< 1.6 \times 10^{-7}$ times the Eddington luminosity, L_{Edd}. Therefore, accretion is highly inefficient in MWC 656, akin to accretion onto black holes in quiescent low-mass X-ray binaries[16], where accretion disks are truncated at ~10^2–10^4 Schwarzschild radii and then behave as an advection dominated accretion flow[17].

In the context of disk instability theory, the very low mass-transfer rates expected for BeXBs (with peak values of ~10^{-11} $M_\odot$ yr^{-1} near periastron) lead to extremely long outburst recurrence periods or even to completely suppressed transient activity[18]. It is the dormant condition of the accretion disk together with the absence of a solid surface reradiating the accretion energy that makes Be–BHs very difficult to detect through X-ray surveys, thus providing an explanation for the missing Be–BH population. This is in stark contrast with the other black-hole high-mass X-ray binary known in the Galaxy, Cygnus X-1, where an X-ray-persistent accretion disk is fed by the powerful wind (~10^{-8} $M_\odot$ yr^{-1}) of an O supergiant star[19].

The detection of a Be–BH is also important for our understanding of BeXB evolution. Whereas the total number of neutron-star BeXBs in the Galaxy depends strongly on the distribution of kick velocities, the number of Be–BHs is very sensitive to the survival probability during the common envelope phase[3]. Modern population synthesis models predict a Galactic number ratio of neutron-star to black-hole BeXBs of 54, for the case of no common envelope survival during the Hertzsprung gap and a Maxwellian distribution of kick velocities with reduced root mean square (r.m.s.) $\sigma = 133$ km s^{-1} (model C in ref. 3). There are currently ~81 BeXBs known in the Galaxy with ~48 pulsating neutron stars[4,20], and thus our discovery of a black-hole companion to MWC 656 is consistent with these model predictions. However, it should be noted that the X-ray spectra of the remaining BeXBs, whenever they are available, also indicate the presence of a neutron star. Further, in stark contrast with the known BeXBs, MWC 656 has been identified through a claimed γ-ray flare (see Methods) and not by its X-ray activity. This seems to imply that the discovery of Be–BHs is observationally biased, in which case common envelope mergers would be less frequent than commonly assumed and/or neutron star kicks would be best described by the radio pulsar birth velocity distribution[3]. Last, it is interesting to note that MWC 656 will probably evolve into a black-hole/neutron-star binary, a potential source of strong gravitational waves and a short γ-ray burst (Methods).

METHODS SUMMARY

The Fe II line was fitted with a two-Gaussian model with Gaussian positions, intensities and separation left as free parameters. The Fe II velocities were obtained from the mean of the Gaussians offset with respect to the rest velocity at 4,583.837 Å. A detailed radial velocity study of the He II profile was performed using the double-Gaussian technique[21], and this shows that the systemic velocity is pushed down in the line core by the S-wave component while the phasing and velocity semiamplitude remain very stable. Therefore, a $+30$ km s^{-1} offset was applied to the He II velocities before fitting all the data points with a double-line orbital model. The S-wave also modulates the He II flux with the orbital period, and its phasing can be interpreted as either enhanced mass transfer during periastron or the visibility of a hotspot in the outer accretion disk. The spectral classification of the star was obtained by direct comparison with a range of templates conveniently broadened, and the best match is provided by the B1.5 III star HD 214993. We used several calibrations available in the literature to constrain the mass of MWC 656 from its spectral type, including evolutionary tracks, a dynamical determination from the detached eclipsing binary V380 Cyg and robust lower limits from dynamical masses of main-sequence stars. The distance is obtained through combining the absolute magnitude of B1.5–2 III stars with the observed brightness of MWC 656, corrected for interstellar reddening. Upper limits to the X-ray flux of MWC 656 are derived from archival ROSAT and Swift pointings, using a neutral hydrogen column density $N_H = 1.4 \times 10^{21}$ cm^{-2} and a photon index $\Gamma = 2.0$, typical of black holes in quiescence. We further discuss the future evolution of MWC 656 and its fate, a possible black-hole/neutron-star binary.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 October; accepted 25 November 2013.

1. Raguzova, N. V. & Lipunov, V. M. The evolutionary evidence for Be/black hole binaries. *Astron. Astrophys.* **349,** 505–510 (1999).
2. Zhang, F., Li, X.-D. & Wang, Z.-R. Where are the Be/black hole binaries? *Astrophys. J.* **603,** 663–668 (2004).
3. Belczynski, K. & Ziolkówski, J. On the apparent lack of Be X-ray binaries with black holes. *Astrophys. J.* **707,** 870–877 (2009).
4. Ziólkowski, J. & Belczynski, K. On the apparent lack of Be X-ray binaries with black holes in the Galaxy and in the Magellanic Clouds. *IAU Symp.* **275,** 329–330 (2011).
5. Casares, J. *et al.* On the binary nature of the γ-ray sources AGL J2241+4454 (=MWC 656) and HESS J0632+057 (=MWC 148). *Mon. Not. R. Astron. Soc.* **421,** 1103–1112 (2012).
6. Williams, S. J. *et al.* The Be star HD 215227: a candidate gamma-ray binary. *Astrophys. J.* **723,** L93–L97 (2010).
7. Reig, P. Be/X-ray binaries. *Astrophys. Space Sci.* **332,** 1–29 (2011).
8. Okazaki, A. T. & Negueruela, I. A natural explanation for periodic X-ray outbursts in Be/X-ray binaries. *Astron. Astrophys.* **377,** 161–174 (2001).
9. Lucarelli, F. *et al.* AGILE detection of the new unidentified gamma-ray source AGL J2241+4454. *Astron. Telegr.* **2761,** 1 (2010).
10. Ballereau, D., Chauville, J. & Zorec, J. High-resolution spectroscopy of southern and equatorial Be stars: flux excess at λ4471 Å. *Astron. Astrophys. Suppl. Ser.* **111,** 423–455 (1995).
11. Smak, J. On the rotational velocities of gaseous rings in close binary systems. *Acta Astronomica* **19,** 155–164 (1969).
12. Etzel, P. SBOP: Spectroscopic Binary Orbit Program (San Diego State Univ., 2004).
13. Hanuschik, R. W. On the structure of Be star disks. *Astron. Astrophys.* **308,** 170–179 (1996).
14. Arias, M. L. *et al.* Fe II emission lines in Be stars. I. Empirical diagnostic of physical conditions in the circumstellar discs. *Astron. Astrophys.* **460,** 821–829 (2006).
15. Peters, G. J., Pewett, T. D., Gies, D. R., Touhami, Y. N. & Grundstrom, E. D. Far-ultraviolet detection of the suspected subdwarf companion to the Be star 59 Cygni. *Astrophys. J.* **765,** 2–9 (2013).
16. Garcia, M. R. *et al.* New evidence for black hole event horizons from Chandra. *Astrophys. J.* **553,** L47–L50 (2001).
17. Esin, A. A., McClintock, J. E. & Narayan, R. Advection-dominated accretion and the spectral states of black hole X-ray binaries: application to Nova Muscae 1991. *Astrophys. J.* **489,** 865–889 (1997).
18. Menou, K., Narayan, R. & Lasota, J.-P. A population of faint nontransient low-mass black hole binaries. *Astrophys. J.* **513,** 811–826 (1999).
19. Coriat, M., Fender, R. P. & Dubus, G. Revisiting a fundamental test of the disc instability model for X-ray binaries. *Mon. Not. R. Astron. Soc.* **424,** 1991–2001 (2012).
20. Linden, T., Valsecchi, F. & Kalogera, V. On the rarity of X-ray binaries with naked helium donors. *Astrophys. J.* **748,** 114–121 (2012).
21. Schneider, D. P. & Young, P. The magnetic maw of 2A 0311-227. *Astrophys. J.* **238,** 946–954 (1980).
22. Smak, J. On the S-wave components of the emission lines in the spectra of cataclysmic variables. *Acta Astronomica* **35,** 351–367 (1985).

Acknowledgements We thank T. Maccarone and P. Charles for comments on the paper. This work made use of the *molly* software package developed by T. R. Marsh. The Liverpool telescope and the Mercator telescope are operated on the island of La Palma by the Liverpool John Moores University and the University of Leuven/Observatory of Geneva, respectively, in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. The Liverpool telescope is funded by the UK Science and Technology Facilities Council. This research was supported by the Spanish MINECO and FEDER under grants AYA2010-18080, AYA2010-21782-C03-01, AYA2010-21967-C05-04/05, AYA2012-39364-C02-01/02, AYA2012-39612-C03-01, FPA2010-22056-C06-02 and SEV2011-0187-01; it was also funded by grant PID

2010119 from the Gobierno de Canarias. J.M.P. acknowledges financial support from ICREA Academia.

Author Contributions J.C. performed the radial velocity analysis of the spectra and wrote the paper. I.N. obtained the Mercator spectrum and contributed to the interpretation of the data. I.R. computed the eccentric orbital fits to the radial velocity curves. M.R. calculated the distance and X-ray luminosity, and contributed to the interpretation of the data. J.M.P. also contributed to the interpretation of the data. A.H. computed the rotational broadening of the star and, together with I.N., performed the spectral calibration of the star. S.S-D. observed the standard stars and reduced the Mercator spectra. J.M.P. and M.R. assisted in writing the section on γ-ray binaries in Methods.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.C. (jorge.casares@iac.es).

METHODS

Spectroscopic observations. Thirty-two 10-min spectra of MWC 656 were obtained between 23 April and 28 July 2011 using the Fibre-fed RObotic Dual-beam Optical Spectrograph (FRODOspec) on the robotic 2.0-m Liverpool telescope at the Observatorio del Roque de Los Muchachos on La Palma (Spain). Full details of these observations are given elsewhere[5]. Four additional FRODOspec spectra were collected on the nights of 28 May and 2–4 June 2012 using an identical instrument configuration. We also used the High Efficiency and Resolution Mercator Echelle Spectrograph (HERMES) on the 1.2-m Mercator telescope to obtain a 15-min spectrum on the night of 26 October 2012. We employed the high-resolution mode in HERMES which yields a resolving power $R = 85{,}000$ across the entire optical range 3,770–9,000 Å. A series of B-type MK standards (where MK refers to the Morgan–Keenan spectral classification) were also observed with HERMES on the nights of 14 June and 9 November 2011 using an identical set-up as for MWC 656. The automatic pipeline products were used for the extraction and calibration of the Liverpool telescope and Mercator telescope data. A full log of the observations is presented in Extended Data Table 1.

Radial velocity analysis. Radial velocities were extracted from the He II 4,686 Å emission by fitting a single Gaussian to the line profile in the 36 Liverpool telescope spectra. A least-squares sine fit to the velocity points yields an orbital period of 59.5 ± 0.6 days (Extended Data Fig. 1a), which is consistent within 1.4σ with the (more accurate) photometric period determination[6]. The difference is also explained by the fact that our radial velocities only cover two full orbital cycles. Consequently, we henceforth adopt 60.37 ± 0.04 days as the true orbital period of the binary. The radial velocities were modelled with an eccentric binary orbit using the code SBOP[12], with the period fixed to 60.37 days. Individual points were weighted proportionally to $1/\sigma^2$, where σ is the radial velocity uncertainty. We adjusted the following orbital parameters: epoch of periastron (T_0), eccentricity (e), argument of the periastron (ω), systemic velocity (γ) and velocity semiamplitude (K). The resulting orbital elements are listed in the first column of Extended Data Table 2, together with their implied fundamental binary parameters. The fitted solution yields maximum velocity at phase 0.06, where phase 0 is arbitrarily set to HJD 2453243.3 or the epoch of maximum brightness in the photometric light curve[6]. This is approximately in antiphase with the radial velocity of the Be star previously determined[5], and therefore the He II emission most probably follows the orbit of the companion star.

To further constrain the orbital motion of the Be star, we measured radial velocities from the Fe II 4,583 Å emission line by fitting a two-Gaussian model to the individual profiles. The Gaussians are set to have identical widths but their intensities and separation are allowed to vary to account for possible profile changes. In any case, these three free parameters are found to be very stable, with differences which are always within 10%. Only in one case was the double-peak separation found to be 13% smaller than the average. This corresponds to one spectrum where the profile appears blurred. For this particular case we decided to fix the Gaussian separation to the average value, 270 km s^{-1}. The radial velocity of the Fe II line was then obtained from the centre of the double Gaussian model relative to the line rest velocity at 4,583.837 Å. We subsequently fitted the radial velocity points with an eccentric orbit model using SBOP after fixing the orbital period to 60.37 days and weighting data points proportionally to $1/\sigma^2$, as before. The best set of fitting parameters are listed in the second column of Extended Data Table 2.

The two independent solutions presented in Extended Data Table 2 give periastron phases T_0 which are consistent at 1.0σ. Also, the arguments of periastron are in antiphase within 1.0σ uncertainties, that is, $\omega_{\rm FeII} = \omega_{\rm HeII} + 180^\circ$. Further, the eccentricities are consistent at the 1.6σ level. These are all strong indications that both radial velocity curves display the reflex motion of two components in a binary system, thus endorsing fitting a combined double orbit to the ensemble of 72 He II and Fe II data points. In any case, the radial velocity amplitudes of the single line orbits are found to be very robust and in excellent agreement with the combined double-lined solution presented in the main text. These are the key parameters constraining the binary mass ratio and thus the nature of the companion to the Be star.

The diagnostic diagram. Extended Data Table 2 shows a 30 km s^{-1} difference between the systemic velocities of the Fe II and He II solutions, which we interpret as due to contamination by the S-wave component seen in the core of the He II profile[22]. To test this, we measured radial velocities from different velocity bands of the He II line profile, using the double-Gaussian technique[21]. To further increase the signal-to-noise ratio, we averaged the 36 Liverpool telescope spectra into 15 phase bins, relative to the photometric ephemeris. Every phase-binned spectrum was convolved with a two-Gaussian passband where the Gaussian separation was varied in the range $a = 200$–600 km s^{-1} in steps of 50 km s^{-1}. Gaussian separations $a < 200$ km s^{-1} produce radial velocities which are corrupted by the variable shape of the double-peaked profile, while for $a > 600$ km s^{-1} the line flux becomes too weak. Each radial velocity curve was subsequently fitted with the expression $V(\varphi) = \gamma + K\sin[2\pi(\varphi - \varphi_0)]$, where φ is the binary phase, and the evolution of the fitting parameters K, γ and φ_0 versus the Gaussian separation a is displayed in Extended Data Fig. 2 as a diagnostic diagram[23]. Note that we prefer to fit a simple sine wave model rather than a full eccentric orbit because the extra fitting parameters (that is, e, ω and periastron phase) become poorly constrained as we approach the noisy wings of the line profile. Given the small orbital eccentricity the fitted sine wave parameters are still meaningful, although they should be taken as a first approximation to the true values because of the model oversimplification.

The high-velocity wings of the He II profile are formed in the inner regions of the accretion disk and are less affected by the core S-wave component. Therefore, they are expected to closely trace the motion of the compact star. Extended Data Fig. 2 shows that, as we move away from the line core, the systemic velocity rises quickly from about -46 km s^{-1} to about -25 km s^{-1}, thus approaching the value of the Fe II solution. This demonstrates that the Fe II velocities provide a more accurate description of the true binary systemic velocity than the centroid of the He II line and, therefore, we decide to apply a $+30$ km s^{-1} offset to the latter. Beyond $a > 500$ km s^{-1} the continuum noise begins to dominate the radial velocities, as indicated by the steeper rise in the control parameter $\sigma(K)/K$ (ref. 23), and thus the fitted parameters are less reliable. The diagram also illustrates that both phasing and velocity semiamplitude are very stable in the interval $a = 200$–500 km s^{-1}, ranging between $\varphi_0 = 0.77$–0.82 and $K = 76$–92 km s^{-1}. Beyond $a > 500$ km s^{-1} K drops below 70 km s^{-1} and hence there is a possibility that the velocity semiamplitude of the compact star was overestimated in Table 1. Note, however, that this would raise the binary mass ratio which would make even stronger the argument for a black-hole companion.

Orbital modulation of the He II flux. Extended Data Fig. 1b shows that the equivalent width of the He II 4,686 Å line is strongly modulated with the orbital period. It peaks at phase ~0.9, which is very close to the maximum in the photometric light curve (that is, phase 0 by convention). On the other hand, the amplitude of the equivalent width modulation is an order of magnitude larger (~40%) than that of the photometric light curve[6,24]. These two arguments imply that the equivalent width variability is driven by true changes in the line flux rather than in the continuum.

Extended Data Fig. 1 also reveals that the maximum equivalent width (phase 0.93 ± 0.04) almost coincides with the peak in the He II radial velocity curve, when the compact star is receding from us at maximum speed. The latter agrees well with the time of maximum visibility of the hotspot or shock region between the gas stream and the accretion disk. Alternatively, the modulation of the He II flux can be interpreted as caused by enhanced mass transfer near periastron, which is constrained to phase 0.01 ± 0.10 by our orbital solution.

Spectral classification, mass and distance to MWC 656. The spectral classification of MWC 656 is complicated by the effects of fast rotation, and by the presence of many (mostly Fe II) emission lines affecting many of the features useful for this purpose. To provide an improved classification, our high-quality Mercator telescope+HERMES spectrum was compared to the spectra of several MK standards taken with the same instrumentation and set-up as for MWC 656. Before this, we measured the rotational broadening ($v\sin i$) in MWC 656 by applying the Fourier technique[25], combined with a goodness-of-fit method[26], to the He I absorption profiles (4,387 Å, 4,471 Å and 4,922 Å) and the Si III line 4,552 Å. This technique allows us to disentangle the different contributions to the line broadening, with the first zeroes of the transformed profile due to rotation. We obtain $v\sin i = 330 \pm 30$ km s^{-1}, with the error reflecting the dispersion of the individual lines, in good agreement with the value previously reported[5]. All our MK standards have very small intrinsic broadenings[27], and thus they were subsequently broadened by 330 km s^{-1} to reproduce the observed broadening in MWC 656. This was done by convolving the MK spectra with a Gray rotational profile[25], using a limb-darkening coefficient $\varepsilon = 0.34$ which is appropriate for the stellar parameters of our target (see below) and the spectral range of interest.

The narrow wings of the Balmer lines definitely indicate that the star is a giant, while the strength of the O II spectrum makes it B2 or earlier, in contrast with a previous classification[6] which reported a B3 IV. The absence of Si IV and He II absorption lines places MWC 656 in the B1–B2 range. As shown in Extended Data Fig. 3, the best match to the overall spectrum is provided by the B1.5 III standard HD 214993, although MWC 656 has rather stronger N II features. Nitrogen enhancement is frequently found in fast rotators[28,29], and understood as a natural product of stellar evolution[30]. Nitrogen enhancement is often accompanied by C depletion. Therefore, we cannot completely rule out the possibility that MWC 656 is a B2 III giant with some C depletion (compare to the spectrum of the MK standard HD 35468 in Extended Data Fig. 3), but the strength of Si III and O II lines, which are the features most sensitive to temperature in this spectral range, strongly supports a B1.5 III classification.

Armed with the spectral classification, we can now set constraints on the mass of the Be star using the several calibrations available in the literature. Based on evolutionary tracks[31], giant stars in the range B1–B2 have mass 12–17 $M_\odot$. On

the other hand, the most precise calibrations are provided by dynamical determinations in detached eclipsing binaries but unfortunately data on B1–B2 giants are very scarce. The closest example is provided by the B1.5 III star in V380 Cyg, where a dynamical mass of $13.1 \pm 0.2\ M_{\odot}$ has been reported[32]. No further analogies are found in a recent exhaustive compilation of detached binaries[33]. In any case, robust lower limits to the mass of our target are provided by dynamical masses of main-sequence stars. For instance, ref. 33 gives $11M_{\odot}$ for B1 V and $9M_{\odot}$ for B2 V, while ref. 34 yields $10M_{\odot}$ for B1.5 V. Taking into account all the above, we support a mass range of 10–16 $M_{\odot}$ for the Be star MWC 656. We also remark that the above mass range might be an underestimate because of the large rotational velocity of MWC 656. This causes the star to appear cooler and slightly less luminous than an object of the same stellar mass rotating at lower speed. This effect could increase our estimated mass by ~10–15%, which would raise the black-hole mass even more[35].

We can also estimate the distance to MWC 656 by comparing the absolute magnitudes of B1.5–B2 III stars with the observed brightness, corrected for interstellar reddening. To estimate the latter, we have constructed the spectral energy distribution (SED) of MWC 656 in the ultraviolet and optical bands[36,37]. The SED was fitted in the range 150–500 nm (where the contribution of the Be disk is marginal) with TLUSTY[38] and FASTWIND[39] models reddened by variable amounts, and the best result is obtained for colour excess $E(B-V) = 0.24$. Adopting an absolute visual magnitude M_V in the range −3.7 to −4.5, based on the calibration of ref. 31, and $V = 8.75 \pm 0.10$ mag (ref. 40), we derive a distance of 2.6 ± 0.6 kpc. Note that in this calculation we have neglected the contribution to the optical light from both the circumstellar Be disk and the black-hole accretion disk. The fact that MWC 656 shows photometric variability modulated with the orbital period argues for some contribution from non-stellar sources to the optical flux. However, the modulation has a semiamplitude of only 0.020 mag, suggesting that these other contributions have a very modest effect[6,24] and, therefore, we can safely conclude that the Be star dominates the observed optical flux. In any case, the large error in our distance estimate encompasses the uncertainty introduced by the small contribution of non-stellar light sources to the observed optical flux. Incidentally, the SED does not show any ultraviolet excess as would be expected if the companion to the Be star were a stripped He core. This gives additional support for the presence of a black hole in MWC 656.

X-ray analysis. Inspection of the HEASARC archives at the position of MWC 656 reveals X-ray observations by ROSAT and Swift. The ROSAT/Position Sensitive Proportional Counters (PSPC) observations (0.1–2.4 keV) were conducted on 7–11 July 1993 (orbital phases 0.63 to 0.70) and consisted of five pointings of ~3–6 ks each. The Swift/X-Ray telescope (XRT) observations (0.3–10 keV) were conducted on 8 March 2011 (orbital phase 0.52) and consisted of two 1-ks pointings separated by around 3 h.

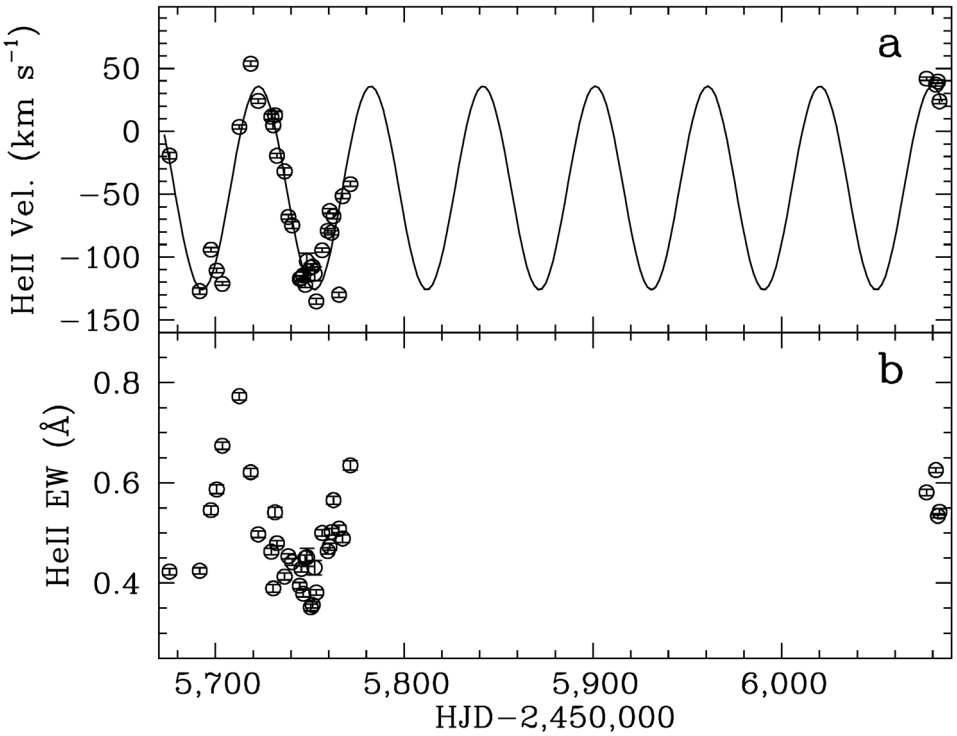
We have analysed both data sets using standard procedures within HEASOFT V 6.12. Assuming a photon index $\Gamma = 2.0$ (typical of black holes in quiescence[41]) and a hydrogen column density $N_H = 1.4 \times 10^{21}$ cm^{-2} (using the obtained $E(B-V)$ and the relation with N_H of ref. 42) the following upper limits (at 90% confidence) are obtained: $F_{0.1-2.4\ \mathrm{keV}} < 1.2 \times 10^{-13}$ erg cm^{-2} s^{-1} for ROSAT/PSPC and $F_{0.3-10\ \mathrm{keV}} < 4.6 \times 10^{-13}$ erg cm^{-2} s^{-1} for Swift/XRT.

γ-ray association. MWC 656 has been proposed as the optical counterpart of the transient GeV γ-ray source AGL J2241+4454 (refs 5, 6). The association is, however, uncertain because AGL J2241+4454 was only detected during a 2-day activity period and it has a position error circle radius of 0.6° (ref. 9). A few binary systems have been detected at GeV and/or TeV energies[43], all showing orbitally modulated γ-ray emission and most of them thought to contain young non-accreting pulsars. In contrast, the accreting black hole Cygnus X-1 showed evidence of TeV emission for less than one day[44], and also two different short transient episodes of GeV emission[45]. The black-hole nature of MWC 656 and its putative flaring γ-ray activity are reminiscent of those of the well-known black hole Cygnus X-1 but, remarkably, the accretion luminosity in MWC 656 is typically lower by more than ~5 orders of magnitude.

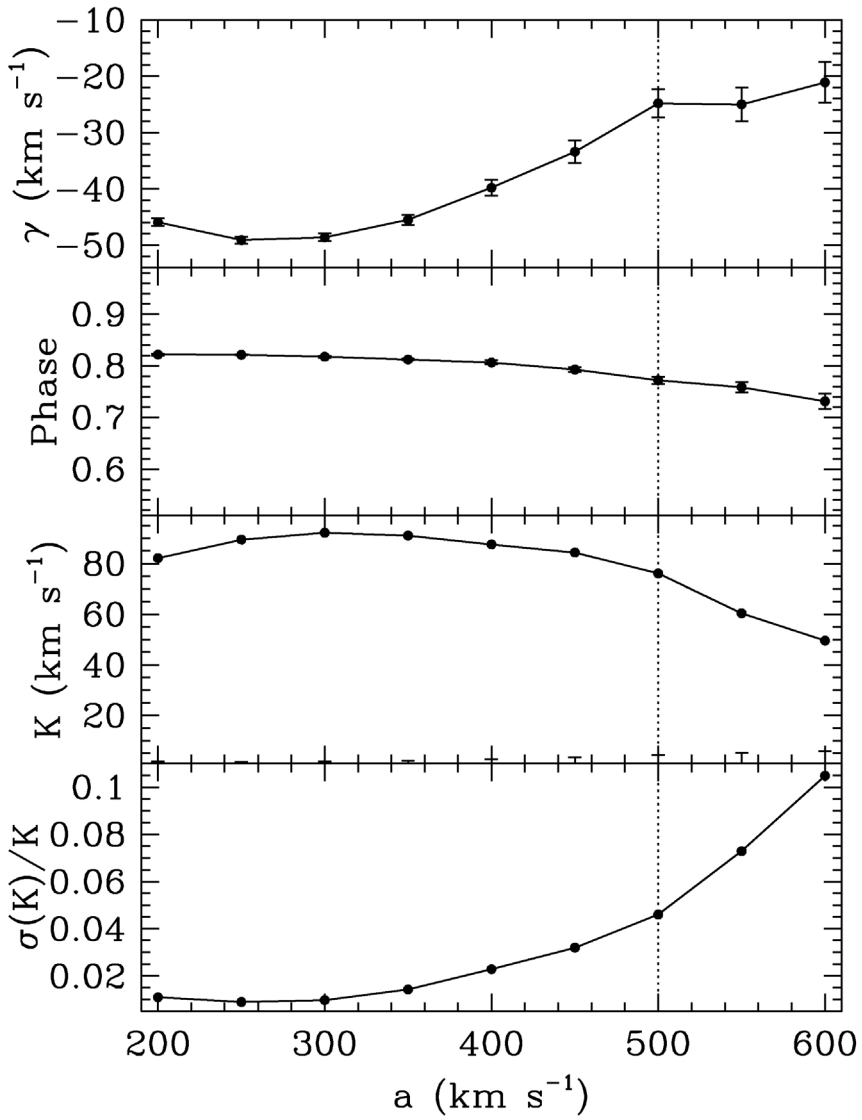
A candidate black-hole/neutron-star progenitor. The future evolution of MWC 656 will probably lead to a black-hole/neutron-star binary[46]. During the red giant phase the $13M_{\odot}$ Be star will expand by several hundred solar radii[47], thus engulfing the black hole. Mass transfer from the expanding Be star onto the black hole will be dynamically unstable and a common envelope will ensue. This is a highly dissipative process which leads to spiral-in of the black hole, efficient circularization of the orbit and the ejection of the Be star envelope. The outcome of the common envelope phase will then be a $2.9M_{\odot}$ He star[48] and the present $\sim 5M_{\odot}$ black-hole companion in a close circular orbit. In the event of a symmetric core collapse, the newly born black-hole/neutron-star binary will remain bound because less than half the total initial mass is expelled in the explosion[49]. In the case of an asymmetric supernova, the binary survival will depend on the magnitude and direction of the kick.

Black-hole/neutron-star binaries, which have not yet been detected, are instrumental in providing fundamental tests of gravitational theories, strong sources of gravitational waves and prime candidates for the production of short γ-ray bursts through coalescence[50–52]. The fate of MWC 656 as a possible black-hole/neutron-star binary is very relevant because it provides tight empirical constraints on detection rates for gravitational wave observatories, such as advanced LIGO/VIRGO[53].

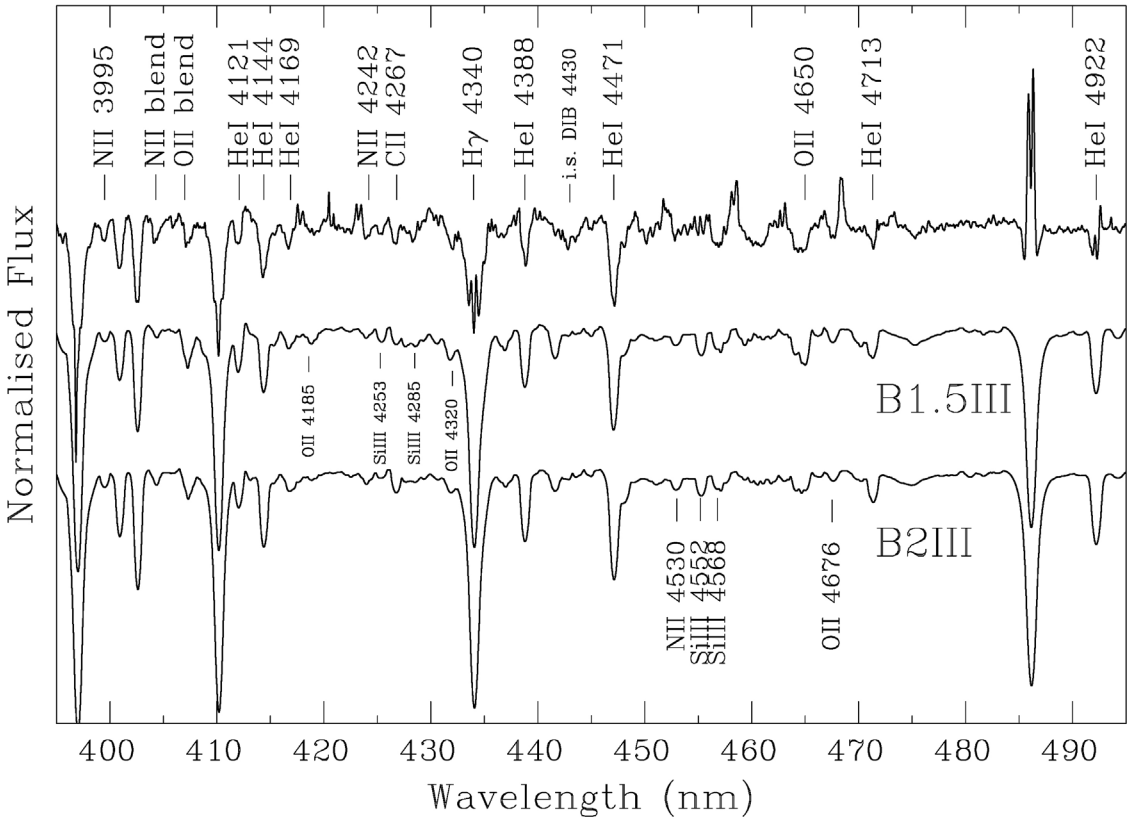
23. Shafter, A. W., Szkody, P. & Thorstensen, J. R. X-ray and optical observations of the ultrashort period dwarf nova SW Ursae Majoris — a likely new DQ Herculis star. *Astrophys. J.* **308,** 765–780 (1986).
24. Paredes-Fortuny, X., Ribó, M., Fors, O. & Núñez, J. Optical photometric monitoring of gamma-ray binaries. *Am. Inst. Phys. Conf. Ser.* **1505,** 390–393 (2012).
25. Gray, D. F. *The Observations and Analysis of Stellar Photospheres* (CUP 20, Wiley-Interscience, 1992).
26. Ryans, R. S. *et al.* Macroturbulent and rotational broadening in the spectra of B-type supergiants. *Astron. Astrophys.* **336,** 577–586 (2002).
27. Abt, H. A., Levato, H. & Grosso, M. Rotational velocities of B stars. *Astrophys. J.* **573,** 359–365 (2002).
28. Howarth, I. D. & Smith, K. C. Rotational mixing in early-type main-sequence stars. *Mon. Not. R. Astron. Soc.* **327,** 353–368 (2001).
29. Walborn, N. R. *et al.* Further results from the Galactic O-star spectroscopic survey: rapidly rotating late ON giants. *Astron. J.* **142,** 150–156 (2011).
30. Przybilla, N., Firnstein, M., Nieva, M. F., Meynet, G. & Maeder, A. Mixing of CNO-cycled matter in massive stars. *Astron. Astrophys.* **517,** A38–A43 (2010).
31. Straižys, V. & Kuriliene, G. Fundamental stellar parameters derived from the evolutionary tracks. *Astrophys. Space Sci.* **80,** 353–368 (1981).
32. Pavlovski, K. *et al.* Chemical evolution of high-mass stars in close binaries — II. The evolved component of the eclipsing binary V380 Cygni. *Mon. Not. R. Astron. Soc.* **400,** 791–804 (2009).
33. Torres, G., Andersen, J. & Giménez, A. Accurate masses and radii of normal stars: modern results and applications. *Astron. Astrophys. Rev.* **18,** 67–126 (2010).
34. Harmanec, P. Stellar masses and radii based on modern binary data. *Bull. Astron. Inst. Czech.* **39,** 329–345 (1988).
35. Negueruela, I. *et al.* Astrophysical parameters of LS 2883 and implications for the PSR B1259-63 gamma-ray binary. *Astrophys. J.* **732,** L11–L15 (2011).
36. Thompson, G. I. *et al. Catalogue of Stellar Ultraviolet Fluxes. A Compilation of Absolute Stellar Fluxes Measured by the Sky Survey Telescope (S2/68) Aboard the ESRO Satellite TD-1* (Science Research Council, UK, 1978).
37. Merrill, P. W. & Burwell, C. G. Supplement to the Mount Wilson catalogue and bibliography of stars of classes B and A whose spectra have bright hydrogen lines. *Astrophys. J.* **98,** 153–184 (1943).
38. Hubeny, I. & Lanz, T. NLTE line blanketed model atmospheres of hot stars. I. Hybrid complete linearization/accelerated lambda iteration method. *Astrophys. J.* **439,** 875–904 (1995).
39. Puls, J. *et al.* Atmospheric NLTE-models for the spectroscopic analysis of blue stars with winds. II. Line-blanketed models. *Astron. Astrophys.* **435,** 669–698 (2005).
40. Nicolet, B. Catalogue of homogeneous data in the UBV photoelectric photometric system. *Astron. Astrophys. Suppl. Ser.* **34,** 1–49 (1978).
41. Plotkin, R. M., Gallo, E. & Jonker, P. G. The X-ray spectral evolution of galactic black hole X-ray binaries toward quiescence. *Astrophys. J.* **773,** 59–74 (2013).
42. Bohlin, R. C., Savage, B. D. & Drake, J. F. A survey of interstellar H I from L-alpha absorption measurements. II. *Astrophys. J.* **224,** 132–134 (1978).
43. Mirabel, I. F. Gamma-ray binaries revealed. *Science* **335,** 175–176 (2012).
44. Albert, J. *et al.* Very high energy gamma-ray radiation from the stellar mass black hole binary Cygnus X-1. *Astrophys. J.* **665,** L51–L54 (2007).
45. Sabatini, S. *et al.* Gamma-ray observations of Cygnus X-1 above 100 MeV in the hard and soft states. *Astrophys. J.* **766,** 83–97 (2013).
46. Lipunov, V. M., Postnov, K. A., Prokhorov, M. E. & Osminkin, E. Yu. Binary radiopulsars with black holes. *Astrophys. J.* **423,** L121–L124 (1994).
47. Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. New grids of stellar models from 0.8 to 120 solar masses at Z = 0.020 and Z = 0.001. *Astron. Astrophys.* **96** (suppl.), 269–331 (1992).
48. Woosley, S. E. & Weaver, T. A. The evolution and explosion of massive stars. II. *Astrophys. J. Suppl. Ser.* **101,** 181–235 (1995).
49. Boersma, J. Mathematical theory of the two-body problem with one of the masses decreasing with time. *Bull. Astron. Inst. Neth.* **15,** 291–301 (1961).
50. Narayan, R., Piran, T. & Shemi, A. Neutron star and black hole binaries in the Galaxy. *Astrophys. J.* **379,** L17–L20 (1991).
51. Portegies Zwart, S. F. & Yungelson, L. R. Formation and evolution of binary neutron stars. *Astron. Astrophys.* **332,** 173–188 (1998).
52. Belczynski, K., Kalogera, V. & Bulik, T. A Comprehensive study of binary compact objects as gravitational wave sources: evolutionary channels, rates, and physical properties. *Astrophys. J.* **572,** 407–431 (2002).
53. Belczynski, K. *et al.* Cyg X-3: a galactic double black hole or black-hole-neutron-star progenitor. *Astrophys. J.* **764,** 96–102 (2013).
54. Lesh, J. R. The kinematics of the Gould Belt: an expanding group? *Astrophys. J.* **17** (suppl.), 371–444 (1969).



Extended Data Figure 1 | Time evolution of the He II 4,686 Å emission line in MWC 656. **a**, Radial velocities obtained from single Gaussian fits to the line profile. The best fitting sine wave, with a period of 59.5 days, is overplotted. Maximum velocity occurs at HJD 2455722.2 or photometric phase 0.06. **b**, Equivalent width (EW) as a function of time. We used the convention of positive equivalent widths for emission lines. Error bars, 1 s.d.



Extended Data Figure 2 | Diagnostic diagram for the He II 4,686 Å line in MWC 656. It has been computed using the double-Gaussian technique with a Gaussian width equal to the instrumental resolution full-width at half-maximum, 55 km s^{-1}. The vertical dotted line indicates the Gaussian separation for which the continuum noise starts to dominate. Error bars, 1 s.d. Panels display the evolution of the sine wave fitting parameters with Gaussian separation a. From top to bottom: the systemic velocity γ, the sine wave φ_0 phase, the velocity semiamplitude K and the control parameter $\sigma(K)/K$.



Extended Data Figure 3 | Classification spectrum of the Be star MWC 656. From top to bottom, spectra of MWC 656 and the MK standards HD 214993 (B1.5 III) and HD 35468 (B2 III) (ref. 54). The standards have been artificially broadened by 330 km s^{-1} to mimic the rotational broadening of MWC 656.

Extended Data Table 1 | Observing log of MWC 656

Date	Telescope+ instrument	Spect. Range Å	# Exp.	Integration (sec.)	Res. $\Delta\lambda/\lambda$
23 Apr – 28 Jul 2011	LT+FRODOspec	3900–5215	32	600	5000
28 May 2012	LT+FRODOspec	3900–5215	1	600	5000
2 – 4 June 2012	LT+FRODOspec	3900–5215	3	600	5000
26 Oct 2012	MT+HERMES	3770–9000	1	900	85000

Extended Data Table 2 | Orbital elements derived from radial velocities of the He II 4,686 Å and Fe II 4,583 Å lines

Parameter	He II λ4686	Fe II λ4583
$P_{\rm orb}$ (days)	60.37 (fixed)	60.37 (fixed)
T_0 (HJD−2,450,000)	3245.3±7.5	3243.1±3.7
e	0.08±0.06	0.24±0.08
ω (deg)	351.7±44.4	164.4±22.1
γ (km s^{-1})	−44.5±3.4	−13.5±1.8
K (km s^{-1})	78.8±4.6	31.0±2.4
$a \sin i$ ($R_\odot$)	93.7±5.4	35.9±2.8
$f(M)$ ($M_\odot$)	3.02±0.53	0.17±0.04
$\sigma_{\rm f}$ (km s^{-1})	19.2	10.3

T_0 is the epoch of periastron, e the orbit eccentricity, ω the longitude of periastron, γ the systemic velocity, K the velocity semiamplitude, a the semimajor axis, i the binary inclination, $f(M)$ the mass function and $\sigma_{\rm f}$ the r.m.s. of the fit.

LETTER

doi:10.1038/nature12806

Nanoparticle solutions as adhesives for gels and biological tissues

Séverine Rose[1], Alexandre Prevoteau[2], Paul Elzière[1], Dominique Hourdet[1], Alba Marcellan[1,2] & Ludwik Leibler[2]

Adhesives are made of polymers[1] because, unlike other materials, polymers ensure good contact between surfaces by covering asperities, and retard the fracture of adhesive joints by dissipating energy under stress[2,3]. But using polymers to 'glue' together polymer gels is difficult, requiring chemical reactions, heating, pH changes, ultraviolet irradiation or an electric field[4–7]. Here we show that strong, rapid adhesion between two hydrogels can be achieved at room temperature by spreading a droplet of a nanoparticle solution on one gel's surface and then bringing the other gel into contact with it. The method relies on the nanoparticles' ability to adsorb onto polymer gels and to act as connectors between polymer chains, and on the ability of polymer chains to reorganize and dissipate energy under stress when adsorbed onto nanoparticles. We demonstrate this approach by pressing together pieces of hydrogels, for approximately 30 seconds, that have the same or different chemical properties or rigidities, using various solutions of silica nanoparticles, to achieve a strong bond. Furthermore, we show that carbon nanotubes and cellulose nanocrystals that do not bond hydrogels together become adhesive when their surface chemistry is modified. To illustrate the promise of the method for biological tissues, we also glued together two cut pieces of calf's liver using a solution of silica nanoparticles. As a rapid, simple and efficient way to assemble gels or tissues, this method is desirable for many emerging technological and medical applications such as microfluidics, actuation, tissue engineering and surgery.

Very often, when brought into contact and pressed together, two pieces of an elastic gel do not stick and gel–gel friction is very low[8]. Incorporating supramolecular or covalent reversible bonds in polymers can produce gels that are self-adhesive[9–11]. However, manipulating self-adhesive gels is not always practical. A solution could be to use supramolecular networks that are able to self-heal when cut into pieces[12]. Dispersing clay particles in polymer solutions or networks yields gels that combine high elasticity and toughness with self-healing capabilities[13,14]. In such gels, the clay particles play the part of reversible crosslinks[15,16]. Nevertheless, self-healing gels cannot meet all the demands of gluing in assembly. At present, gluing gels or biological tissues requires complex methods, often involving *in situ* polymerization[6,17] or electrophoretic transport of polymers to the interface[5,7]. Therefore, it is desirable to find new ways of gluing that are practical and adaptable. Inspired by the physics of polymer adsorption, we propose to use nanoparticle solutions as an adhesive (Fig. 1).

The design principle is simple: to act as an efficient glue, the particles must be adsorbed onto the gel surface. The surface of the particles must therefore exhibit an affinity with network chains, that is, the free energy gain, ε, resulting from the adsorption of a monomer to the surface of a particle should be comparable with the thermal energy, kT (ref. 18). Consider a particle adsorbed onto a surface of a gel having swelling degree Q, the swelling degree being defined as the ratio of the volume of the swollen gel to the volume of the dry network. The number of adsorbed monomers per network strand, n, can be estimated using theoretical tools analogous to those developed for adsorption of polymer solutions. We expect that there are many monomers adsorbed per strand, $n \gg 1$, and that for nanoparticles with diameters comparable with the network mesh size, several different network strands are adsorbed to the same particle[18]. Thus, in the adhesive layer, nanoparticles act as connectors between gel pieces and gel-chains act as bridges between different particles (Fig. 1a). When the adhesive junction is strained, adsorbed chains are under tension (for example, the red chain in Fig. 1b) and some adsorbed monomers detach from the particle surface and relax the tension. Nanoparticles retard failure and ensure good adhesion because the energy dissipated during the stress-induced desorption process is much greater than ε. Indeed, detaching only one

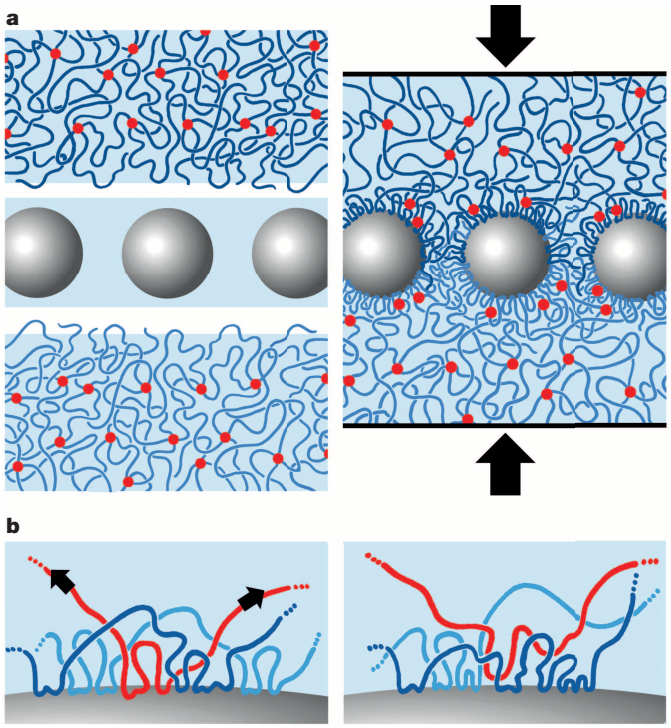


Figure 1 | Gluing gels by nanoparticle solutions. **a**, Schematic illustration of the concept of gluing swollen polymer networks together using particles. The nanoparticle diameter is comparable with the gel network mesh size. Network chains are adsorbed on nanoparticles and anchor particles to gel pieces. Particles act as connectors between gel surfaces. Adsorbed chains also form bridges between particles. The black arrows indicate the pressure applied to squeeze the gel layers together. **b**, Particle adsorption is irreversible because particles are anchored to the gel networks by numerous attachments (red, light- and dark-blue strands). At equilibrium or under tension (indicated by black arrows) a monomer that detaches from a particle surface (red strand) can be replaced by a monomer belonging to the same or a different network strand (shown here as a dark-blue strand). Such exchange processes and rearrangements allow for large deformations and energy dissipation under stress.

[1]Physico-chimie des Polymères et Milieux Dispersés (UMR7615 UPMC/ESPCI/CNRS), Université Pierre et Marie Curie, École Supérieure de Physique et de Chimie industrielles, 75005 Paris, France.
[2]Matière Molle et Chimie (UMR 7167 CNRS/ESPCI), École Supérieure de Physique et de Chimie Industrielles, 75005 Paris, France.

monomer releases the tension in the whole segment of N/n monomers (N is the number of monomers per network strand or, equivalently, the number of monomers between two crosslinks). This mechanism resembles one proposed to explain rubber resistance to fracture[2], but here desorption rather than chain breaking is responsible for stress release and energy dissipation. Furthermore, in contrast to chain breaking, which is irreversible, in adsorbed layers there is a constant traffic of monomers between the adsorbed and desorbed states[19] and a neighbouring strand (for example, the dark-blue chain in Fig. 1b) may adsorb and replace the detached link. These exchange processes enable the adhesive junction to sustain large deformation and to dissipate energy[20], yielding an efficient resistance to interfacial fracture propagation and a strong adhesion.

Gluing by nanoparticles may seem paradoxical given that powdering surfaces with micrometre-sized particles such as talc provides a standard means of preventing self-adhesion. Yet Gent and co-workers have shown that when the particle–elastomer interactions are suitably controlled, powdering self-adhesive elastomers can increase the self-adhesion energy by a factor of two[21,22]. We go a step further to show that systems mostly composed of solvent and which do not adhere to themselves can be bonded by particle solutions.

To demonstrate the concept and test the importance of gel chain adsorption onto particles, we synthesized two hydrogels: S0.1, made of poly(dimethylacrylamide) (PDMA) and A0.1, made of polyacrylamide. Both gels have the same crosslinking density and similar swelling degree and do not adhere to themselves (Fig. 2a and Extended Data Table 1). Polyacrylamide does not adsorb onto silica[23] whereas poly (dimethylacrylamide) adsorbs readily[24]. When a 15-μl drop of TM-50 silica suspension was spread on the PDMA gel surface and another PDMA piece was pressed to form a lap junction, a strong adhesion was observed after a few seconds of contact (Fig. 2b, c and Supplementary Movie 1). In contrast, when using polyacrylamide gels, even when we pressed very firmly and for a very long time, we could not create lap junctions that held under their own weight, thus confirming that the lack of gel chains being adsorbed onto nanoparticles prevents gluing.

Joints made of S0.1 gel were stronger than the gel itself and failure occurred outside the bonding junction. Only when the overlap length was made comparable to the ribbon thickness did interfacial failure by peeling occur (Fig. 2d and Extended Data Figs 1 and 2). From the measured failure force, the adhesion energy[1,25], G_{adh}, could be estimated to be $6.6 \pm 1.6\ J\ m^{-2}$ for short and thin joints, and $6.2 \pm 1.4\ J\ m^{-2}$ for narrow and thick joints. Strong adhesion was also achieved using silica particles of different sizes (Fig. 2e and Extended Data Fig. 3). Adhesion strength increased when particle size was increased, although it should be noted that changing the size of particles implies variations of the surface chemistry as well.

Indeed, adapting particle surface chemistry provides a powerful tool with which to produce adhesion by improving suspension stability and by promoting specific interactions such as hydrogen bonding that strengthen the particle adsorption to the gel surface. Thus, grafting thymine to carbon nanotubes produced adhesion. Similarly, CNC1 cellulose nanocrystals bearing sulphate groups yield an adhesion strength comparable with that obtained with nanosilica, whereas CNC2 particles bearing hydroxyl groups only were useless as a glue (Fig. 2e).

Gluing strength also depends on the properties of the gel. In tightly crosslinked gels, strands are more constrained and there is a higher entropy penalty for adsorption[26]. In addition, the energy dissipation mechanisms discussed above are less efficient for short strands (when N is small). The adhesion energy G_{adh} was therefore weaker for more rigid gels (Fig. 2f). But, more rigid gels are also more brittle, as shown by single-edge notched tensile test results. Hence, in practice, even for rigid gels, gluing by nanoparticles is sufficiently strong to allow the design of joints that withstand tensile stresses without adhesive failure. It is also possible to glue gels with very different rigidities strongly together (Extended Data Fig. 4).

Gluing by nanoparticle solutions is not limited to hydrogels that are in the as-synthesized state. Dehydration of the gels before gluing is expected to increase adhesive strength, whereas gel swelling should have the opposite effect. Indeed, in a more swollen state, the strands are more stretched and adsorb less readily. However, by choosing particles (AL-30) of the ideal size and gel affinity, we were able to glue together completely swollen S0.1 gels containing as much as 97.6 v/v% of water. Adhesion energy was as high as $1.6 \pm 0.6\ J\ m^{-2}$ (Extended Data Fig. 5).

The design principle ensures that adhesion remains when the joint is immersed in excess solvent and swells. Indeed, particles once adsorbed stay strongly anchored to the gel surfaces, the probability of total desorption of a gel strand being exponentially small[18,19]: $\exp(-n\varepsilon/kT)$. Scanning electron microscopy confirmed that adsorbed nanoparticles cover the surface densely, even after soaking the gel in pure water and washing the surface several times (Fig. 3a). When a lap joint made of S0.1 hydrogels glued with TM-50 nanoparticles was immersed in water it withstood a fivefold volume increase without failure (Fig. 3b). The adhesion energy as measured by a lap-shear test at maximum swelling was $1.8 \pm 0.5\ J\ m^{-2}$, that is, 3.5 times lower than in the as-synthesized

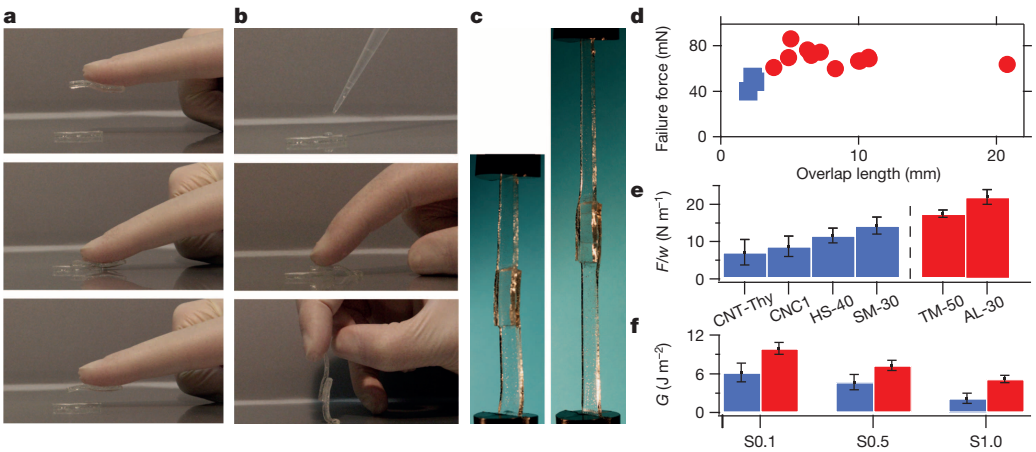


Figure 2 | Lap-shear adhesion tests. **a**, Lightly crosslinked S0.1 gels stick to the table surface and to gloves but they do not stick to themselves. **b**, By spreading a drop of TM-50 silica solution on the gel surface, two gel pieces are glued together after being brought into contact for few seconds. **c**, The glued lap joint is able to sustain large deformations. **d**, The failure force measured by the lap-shear adhesion test for lap joints of various overlap length. Red circles indicate fracture outside the joint; blue squares indicate interfacial failure by peeling. **e**, Failure force, F, normalized by the width of the joint, w, for lap joints glued using solutions of various particles. Red bars indicate fracture outside the joint; blue bars indicate failure by peeling (mean; error bars are s.d.). **f**, Adhesion energy G_{adh} (in blue) measured by the lap-shear test and fracture energy G_c (in red) measured by the single-edge notch tensile test for PDMA gels of various crosslinking densities (mean; errors bars are s.d.).

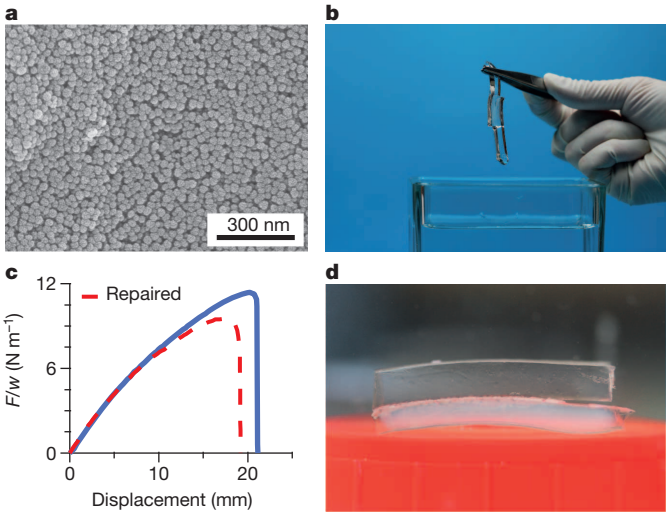


Figure 3 | Water-resistant and self-repairing glue. **a**, Scanning electron micrograph showing that the TM-50 silica layer adsorbed on a S0.1 gel surface was not removed by multiple washing and soaking of the gel in water for several days. **b**, After reaching the maximum swelling and being immersed in deionized water for four weeks, the PDMA S0.1/S0.1 assembly glued together using TM-50 silica held well. **c**, Lap-shear test showing that after adhesive interfacial failure by peeling (blue line), the S0.1/S0.1 joint glued using HS-40 silica solution was repaired by bringing ribbons back into contact with finger pressure for about 30 s (red line). **d**, Gelatine/PDMA S0.1 junction glued with TM-50 silica after immersion in water for one week to reach maximum equilibrium swelling.

state (Extended Data Fig. 5 and Supplementary Video 2). In the fully swollen state the adhesion strength is weaker because detaching adsorbed chains is easier as a result of the higher swelling tension the adsorbed chains are already under. Moreover, once detached, a swollen strand is less prone to adsorb and strand–strand exchange dissipation processes are hindered. The adhesive joint made by adsorbing nanoparticles (Fig.1) ressembles a slice of a nanocomposite gel. Hence, a nanocomposite PDMA/silica gel can be considered to be a bulk model of the adhesive joint. The highly reduced deformation energy dissipation observed in a fully swollen nanocomposite gel seems to confirm the above interpretation of decrease of adhesion when the gel is swollen (Extended Data Fig. 5).

The strong irreversible anchoring of once-adsorbed particles suggests the attractive possibility of self-repairing or re-positioning adhesive joints. Figure 3c shows an example of a joint, which was peeled, but that can recover its initial strength when the ribbons are brought back into contact and pressed with fingers for a few seconds, without any need to re-apply the glue.

Particle solutions offer a simple method of gluing gels of different chemical nature, provided that the particle surface chemistry is properly adjusted to allow adsorption on both gels. For example, using the TM-50 silica solution, a robust assembly of S0.1 and gelatine was achieved (Extended Data Fig. 4). In many applications (such as actuation), gluing gels of different rather than identical chemical nature together presents advantages, such as the possibility of assembling gels with different rigidity, but similar equilibrium swelling. Such assemblies can withstand swelling in excess water (Fig. 3d). In contrast, the swelling of a glued assembly of chemically identical gels with mismatched swelling capacities can lead to high, heterogeneous osmotic stresses near the interface, resulting in slow interfacial failure (Extended Data Fig. 4).

Soft biological tissues, although they are incomparably more complex, both mechanically and osmotically, resemble gels in many respects. To test the gluing potential of nanoparticle solutions we cut two ribbons 45 mm × 18 mm × 3 mm of calf liver. Cut pieces do not adhere to each other and cannot be glued by water at pH 9. We spread 60 µl of silica TM-50 solution on the cut surface (without any pre-treatment or special drying) to make a lap joint with overlap length $l = 20$ mm.

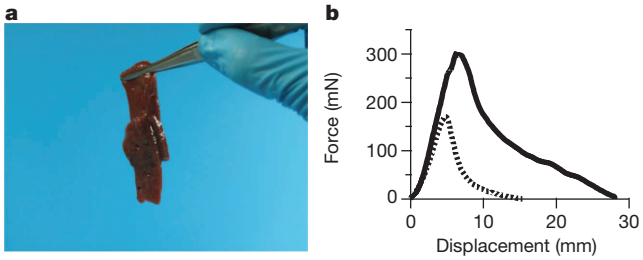


Figure 4 | Gluing biological tissues. **a**, Two pieces cut from calf liver were glued together by spreading TM-50 nanosilica solution between them and then pressing them with a finger. After 30 s of contact the assembly was manipulable and the bond held well. **b**, Normalized force–displacement curves for lap joints made of ribbons cut from calf liver and glued by spreading TM-50 silica solution between them and then pressing ribbons with a finger for 30 s. The ribbons were cut with a scalpel blade and no treatment was applied to liver surfaces before gluing. Results for two livers are presented.

After being pressed for 30 s with a finger, the lap joint held strongly and could be manipulated with ease (Fig. 4a). Lap-shear adhesion tests on two livers yield adhesion energy $G_{adh} \approx 25 \pm 5\ J\ m^{-2}$ and $G_{adh} \approx 6 \pm 1.6\ J\ m^{-2}$ (Fig. 4b and Supplementary Video 3).

The results suggest that nanoparticle solutions provide a simple way of assembling synthetic and biological hydrogels as well as biological tissues without affecting substantially the rigidity or permeability of the assembly. Powerful methods exist to tune and control the surface chemistry of inorganic particles and latexes to achieve optimal particle adsorption and bonding. The possibility of self-repairing and re-positioning peeled adhesive joints is an additional boon. Given the importance of wet adhesion in biomedicine and biotechnology as well as in more traditional coating and material technologies, our results suggest ways to develop new applications by simply assembling many kinds of chemically and mechanically mismatched tissues and gels.

METHODS SUMMARY

Silica Ludox TM-50, HS-40 and SM-30 water solutions with, respectively, concentrations of 52 wt%, 40 wt% and 30 wt% at pH 9, pH 9.5 and pH 10, SiO_2/Na_2O ratios of 200–250, 89–101 and 45–56, and radii of about 15 nm, 9 nm and 5 nm, were purchased from Aldrich and used as received. Stöber silica particles[27], AL-30, with radius 50 nm were synthesized and dissolved in water at 30 wt% (pH = 8.5). Multi-wall carbon nanotubes were supplied by Arkema (Graphistrength C100) and purified with sulphuric acid. Thymine-grafted carbon nanotube particles (CNT-Thy) were synthesized using the method of ref. 28. Cellulose nanocrystals CNC1 bearing sulphate and hydroxyl groups were prepared using the method of ref. 29. Cellulose nanocrystals CNC2 were prepared using the same method, but replacing sulphuric with hydrochloric acid. The suspensions were diluted to the desired concentrations (0.5 wt% and 3 wt% for carbon nanotubes and cellulose nanocrystals, respectively) were sonicated for 30 min just before use.

PDMA and PDMA/silica nanocomposite gels were prepared using the method of ref. 15. The polyacrylamide A0.1 hydrogel was prepared by *in situ* free radical polymerization of acrylamide using thermal dissociation of potassium persulphate (KPS) as initiator, at 80 °C. *N,N′*-methylenebisacrylamide (MBA) was used as the crosslinker; the MBA/dimethylacrylamide ratios were 0.1 mol.%, 0.5 mol.%, 1 mol.% and 1.5 mol.%, for samples S0.1, S0.5, S1.0 and S1.5, respectively, and MBA/acrylamide was 0.1 mol.% for the A0.1 gel. At the preparation state, the gel matrix hydration was fixed at 87.7 wt%. Gelatine (Technical 1, VVR) gels were prepared at 23 wt% in aqueous solutions.

Scanning electron micrographs were obtained using a Field Emission scanning electron microscope (Hitachi SU-70). Lap-shear and mechanical tests were performed on an Instron 5565 machine. Single lap-shear geometry was used for adhesion tests. Gluing was achieved by applying a contact pressure of 10 kPa for 30 s. Fracture energy G_c was measured with the single-edge notch tensile test[30].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 May; accepted 24 October 2013.
Published online 11 December 2013.

1. Kendall, K. *Molecular Adhesion and Its Applications* (Plenum, 2001).

2. Lake, G. J. & Thomas, A. G. The strength of highly elastic materials. *Proc. R. Soc.* **300,** 108–119 (1967).
3. de Gennes, P. G. Soft adhesives. *Langmuir* **12,** 4497–4500 (1996).
4. Sahlin, J. J. & Peppas, N. A. Enhanced hydrogel adhesion by polymer interdiffusion: use of linear poly(ethylene glycol) as an adhesion promoter. *J. Biomater. Sci. Polym. Edn* **8,** 421–436 (1997).
5. Tamagawa, H. & Takahashi, Y. Adhesion force behavior between two gels attached with an electrolytic polymer liquid. *Mater. Chem. Phys.* **107,** 164–170 (2008).
6. Saito, J. *et al.* Robust bonding and one-step facile synthesis of tough hydrogels with desirable shape by virtue of the double network structure. *Polymer Chem.* **2,** 575–580 (2011).
7. Techawanitchai, P. *et al.* Photo-switchable control of pH-responsive actuators via pH jump reaction. *Soft Matter* **8,** 2844–2851 (2012).
8. Gong, J. P. Friction and lubrication of hydrogels—its richness and complexity. *Soft Matter* **2,** 544–552 (2006).
9. Reutenauer, P., Buhler, E., Boul, P. J., Candau, S. J. & Lehn, J.-M. Room temperature dynamic polymers based on Diels-Alder chemistry. *Chem. Eur. J.* **15,** 1893–1900 (2009).
10. Nicolaÿ, R., Kamada, J., Van Wassen, A. & Matyjaszewski, K. Responsive gels based on a dynamic covalent trithiocarbonate cross-linker. *Macromolecules* **43,** 4355–4361 (2010).
11. Harada, A., Kobayashi, R., Takashima, Y., Hashidzume, A. & Yamaguchi, H. Macroscopic self-assembly through molecular recognition. *Nature Chem.* **3,** 34–37 (2011).
12. Cordier, P., Tournilhac, F., Soulié-Ziakovic, C. & Leibler, L. Self-healing and thermoreversible rubber from supramolecular assembly. *Nature* **451,** 977–980 (2008).
13. Wang, Q. *et al.* High-water-content mouldable hydrogels by mixing clay and a dendritic molecular binder. *Nature* **463,** 339–343 (2010).
14. Haraguchi, K., Uyama, K. & Tanimoto, H. Self-healing in nanocomposite hydrogels. *Macromol. Rapid Commun.* **32,** 1253–1258 (2011).
15. Carlsson, L., Rose, S., Hourdet, D. & Marcellan, A. Nano-hybrid self-crosslinked PDMA/silica hydrogels. *Soft Matter* **6,** 3619–3631 (2010).
16. Gaharwar, A. K., Rivera, C. P., Wu, C.-J. & Schmidt, G. Transparent, elastomeric and tough hydrogels from poly(ethylene glycol) and silicate nanoparticles. *Acta Biomater.* **7,** 4139–4148 (2011).
17. Duarte, A. P., Coelho, J. F., Bordado, J. C., Cidade, M. T. & Gil, M. H. Surgical adhesives: systematic review of the main types and development forecast. *Prog. Polym. Sci.* **37,** 1031–1050 (2012).
18. Netz, R. R. & Andelman, D. Neutral and charged polymers at interfaces. *Phys. Rep.* **380,** 1–95 (2003).
19. Santore, M. M. Dynamics in adsorbed homopolymer layers: understanding complexity from simple starting points. *Curr. Opin. Colloid Interf. Sci.* **10,** 176–183 (2005).
20. Montarnal, D., Capelot, M., Tournilhac, F. & Leibler, L. Silica-like malleable materials from permanent organic networks. *Science* **334,** 965–968 (2011).
21. Gent, A. N., Hamed, G. R. & Hung, W. J. Adhesion of elastomer layers to an interposed layer of filler particles. *J. Adhes.* **79,** 905–913 (2003).
22. Nah, C., Jose, J., Ahn, J. H., Lee, Y. S. & Gent, A. N. Adhesion of carbon black to elastomers. *Polym. Test.* **31,** 248–253 (2012).
23. Griot, O. & Kitchener, J. A. Role of surface silanol groups in the flocculation of silica suspensions by polyacrylamide. Part 1—Chemistry of the adsorption process. *Trans. Faraday Soc.* **61,** 1026–1031 (1965).
24. Hourdet, D. & Petit, L. Hybrid hydrogels: macromolecular assemblies through inorganic cross-linkers. *Macromol. Symp.* **291–292,** 144–158 (2010).
25. Kendall, K. Cracking of short lap joints. *J. Adhes.* **7,** 137–140 (1975).
26. Johner, A. & Joanny, J.-F. Adsorption of polymeric brushes: bridging. *J. Chem. Phys.* **96,** 6257 (1992).
27. Stöber, W., Fink, A. & Bohn, E. Controlled growth of monodisperse silica spheres in the micron size range. *J. Colloid Interf. Sci.* **26,** 62–69 (1968).
28. Prevoteau, A., Soulié-Ziakovic, C. & Leibler, L. Universally dispersible carbon nanotubes. *J. Am. Chem. Soc.* **134,** 19961–19964 (2012).
29. Bondeson, D., Mathew, A. & Oksman, K. Optimization of the isolation of nanocrystals from microcrystalline cellulose by acid hydrolysis. *Cellulose* **13,** 171–180 (2006).
30. Greensmith, H. W. Rupture of rubber. X. The change in stored energy on making a small cut in a test piece held in simple extension. *J. Appl. Polym. Sci.* **7,** 993–1002 (1963).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Montero and L. Olanier for technical assistance with electron microscopy and tensile test equipment, respectively. We also thank L. Alison and L. Nègre for help with experiments and A. Legrand for synthesis of silica AL-30 particles. We thank A. Johner for discussions on adsorption and M. Cloitre, J. Lewiner, A. Maggs, R. Nicolaÿ and F. Tournilhac for encouragements and discussions. S.R. and A.P. acknowledge PhD fellowship funding from ED397, UPMC, Paris France. The financial support of the CNRS, the ESPCI and the UPMC is acknowledged.

Author Contributions S.R. synthesized samples, designed and performed experiments, analysed data and discussed results, A.P. synthesized carbon nanotube and CNT-Thy samples and performed experiments, P.E. performed experiments and analysed data, D.H. advised on gel synthesis, A.M. conceived the project, designed and performed experiments, analysed and discussed results and wrote the manuscript, L.L. initiated and conceived the project, designed and performed experiments, interpreted results, and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.L. (ludwik.leibler@espci.fr) or A.M. (alba.marcellan@espci.fr).

METHODS

Particles. Silica Ludox TM-50, HS-40 and SM-30 water solutions with, respectively, concentrations of 52 wt%, 40 wt% and 30 wt% at pH 9, pH 9.5 and pH 10, SiO_2/Na_2O ratios of 200–250, 89–101 and 45–56, and radii of about 15 nm, 9 nm and 5 nm, were purchased from Aldrich and used as received. Silica particles AL-30 with radius about 50 nm were synthesized using the method of ref. 27 and dissolved in water at 30 wt% (pH = 8.5).

Multi-wall carbon nanotubes were supplied by Arkema (Graphistrength C100) and purified with sulphuric acid. Thymine-grafted carbon nanotube particles (CNT-Thy) were synthesized using the method of ref. 28.

Microgranular cellulose (cotton) was purchased from Aldrich. Cellulose nanocrystals CNC1 bearing sulphate and hydroxyl groups were prepared using the method of ref. 29. Cellulose nanocrystals CNC2 were prepared using the same method, but replacing sulphuric with hydrochloric acid. The suspensions were diluted to the desired concentration (0.5 wt% and 3 wt% for carbon nanotubes and cellulose nanocrystals, respectively) were sonicated for 30 min just before use.

Gels and calf liver. PDMA gels were prepared following the method of ref. 15. *N,N′*-methylene bisacrylamide (MBA) was used as the crosslinker and the MBA/dimethylacrylamide ratio was 0.1 mol.%, 0.5 mol.%, 1.0 mol.% and 1.5 mol.% for samples S0.1, S0.5, S1.0 and S1.5, respectively (Extended Data Table 1). Potassium persulphate (KPS) and *N,N,N′,N′*-tetramethylethylenediamine (TEMED) were used as redox initiators. At the preparation state, the gel matrix hydration was fixed at 87.7 wt%. To avoid network defects that lead to a weak self-adhesion of gels it is important to conduct the synthesis under nitrogen conditions. The PDMA/silica nanocomposite gel containing 21 v/v% of TM-50 silica was prepared following the method of ref. 15.

Polyacrylamide hydrogels were prepared by *in situ* free radical polymerization of acrylamide using thermal dissociation of KPS as initiator, at 80 °C. The MBA/acrylamide ratio was 0.1 mol.% for the A0.1 gel. The A0.1 gel hydration at preparation was fixed at 87.7 wt%. Two aqueous solutions were prepared: KPS at 4.5 wt% and MBA at 1.2 wt%. The samples were prepared by first dissolving MBA and acrylamide at 25 °C in water. The KPS solution was then added and the homogeneous solution was purged with nitrogen for 15 min under magnetic stirring. The mixture was finally transferred, under a nitrogen atmosphere, into laboratory-made moulds previously sealed and put under a nitrogen atmosphere. The sealed moulds were placed in an oven at 80 °C for 24 h.

Gelatine (Technical 1, VVR) gels were prepared by dissolving the gelatine powder in deionized water under stirring at 50 °C for 2 h. The gelatine concentration was 23 wt.%. The mixture was then poured into moulds and left at room temperature for 30 min. Moulds were stored at 6 °C for two days before testing and allowed to come to room temperature for 1 h before testing.

Swelling experiments were performed in deionized water and the swollen weight of the gel was recorded over four days to ensure that the gel was at equilibrium. Gels of intermediate swelling degree (Extended Data Fig. 5a) were obtained by immersing as-synthesized gels in deionized water for some time (30 min, 1 h, 2 h and 4 h, respectively) followed by a re-equilibration of samples over 24 h. The amount of dry polymer was estimated from the amount of monomer, assuming 100% conversion.

Fresh calf livers were purchased from M. Bonjean, rue Montagne Ste Geneviève, Paris and used as received. Samples for tensile and lap-shear adhesion tests were cut with a scalpel to the desired dimensions. No pre-treatment or special drying was applied to samples or surfaces before the tensile and adhesion tests.

Tensile tests. Tensile tests were performed on a tensile Instron machine (model 5565) equipped with a 10 N load cell and with a video extensometer, which follows the local displacements of two spots. The experiments were performed at room temperature at strain rate of 0.06 s^{-1}. For PDMA and polyacrylamide, gel ribbons with dimensions of 40 mm × 5 mm × 2 mm were used. For gelatine, to avoid systematic failure in the vicinity of the clamps, samples were cut using a dog-bone-shaped die cutter following the ISO4661-1 standard with the reduced section of the samples having dimensions of of 25 mm × 4 mm × 2 mm. For calf livers, tensile tests were performed on ribbons with dimensions 45 mm × 18 mm × 3 mm. The elastic moduli of the two livers shown in Fig. 4b were respectively 15.0 ± 1.7 kPa and 12 ± 1.5 kPa.

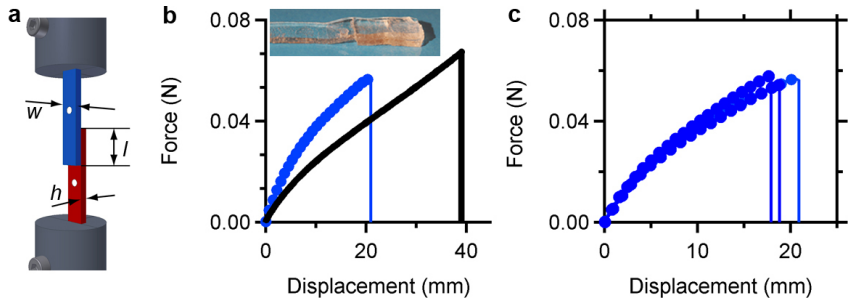
Adhesion tests. Lap-shear tests were performed on an Instron 5565 machine equipped with a 10 N load cell at a speed of 150 mm min^{-1}. We used single lap-shear geometry (Extended Data Fig. 1a). Displacement was measured by a video extensometer that followed two markers (white dots), which were placed at a distance of 5 mm from the edge of the lap joint unless otherwise stated in the text. The total length of assembled ribbons was 40 mm. Unless otherwise stated, gluing was achieved by spreading a nanoparticle solution droplet of volume proportional to the overlap surface with a volume of 0.3 µl per square millimetre. The contact pressure was 10 kPa and contact time 30 s.

In Fig. 2c the overlap length and width are 20 mm and 10 mm, respectively, and the ribbon thickness is $h = 2$ mm. The data of Fig. 2d were obtained for ribbons with $w = 5$ mm and $h = 2$ mm. The data of Fig. 2e were obtained for lap joints with overlap length of $l = 5$ mm made of ribbons with $w = 5$ mm and $h = 2$ mm. The data of Extended Data Figs 1b, c, 4a and 5b were obtained for lap joints with overlap length of 10 mm made of ribbons with $w = 5$ mm and $h = 2$ mm.

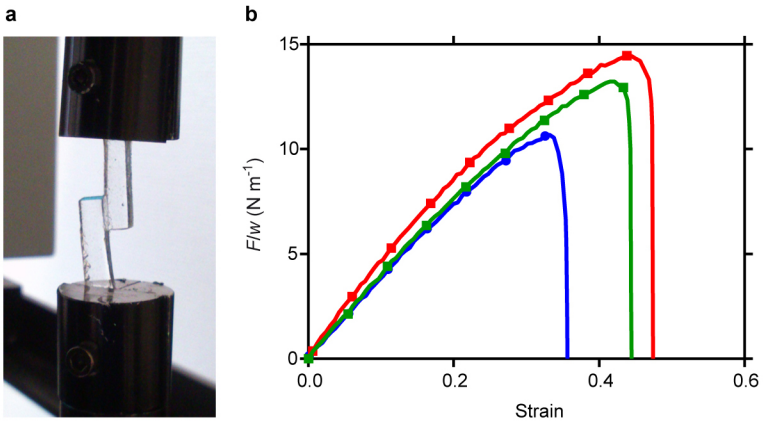
For the lap-shear test, when failure occurred by interfacial peeling, we evaluated the adhesion energy from the measured adhesive failure force F using the expression for short lap joints[26], $G_{adh} = 3(F/w)^2/(2Eh)$, where w and h denote, respectively, the width and thickness of the ribbon, and E is the tensile modulus. For PDMA S0.1 gels glued using TM-50 silica particles we found G_{adh} to be 6.6 ± 1.6 J m^{-2} and 6.2 ± 1.4 J m^{-2}, respectively, for short and narrow ($l = 2$ mm, $w = 5$ mm, $h = 2$ mm) and thick ($l = 5$ mm, $w = 2$ mm, $h = 5$ mm) joints. For S0.5 and S1.0 gels (Fig. 2f) the lap-joint dimensions were $l = 5$ mm, $w = 5$ mm, $h = 2$ mm.

Self-repair tests. The data of Fig. 3c were made of PDMA S0.1 ribbons of $w = 5$ mm and $h = 2$ mm glued by spreading a 6 µl droplet of HS-40 silica solution between them. The initial length of the assembly before the deformation is about 28 mm. The overlap length was 5 mm. After adhesive interfacial failure by peeling, the joint was repaired by bringing ribbons back to contact and pressing with fingers for about 12 s and the lap-shear test was performed again to test the strength of the repaired junction. Similar tests were carried out on gels glued together using other particles (for example, CNC1) and the ability to self-repair was always observed.

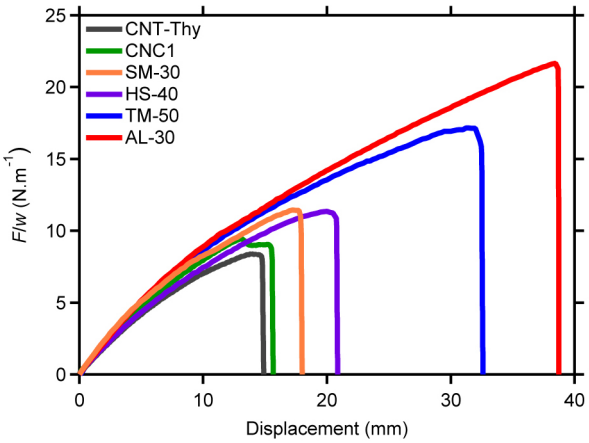
Fracture tests. The fracture energy data of Fig. 2f were obtained using the single-edge notch geometry. A cut was made in the centre of the samples using a blade. Each notch was measured by optical microscopy to determine its exact length a, of approximately 1 mm. The same procedure as the one used for tensile tests was performed: the strain rate was fixed to be 0.06 s^{-1}, and the force and the displacement data were recorded. The fracture energy was calculated using the following expression: $G_c = 6aW/(\lambda_c)^{1/2}$. Here W is the strain energy density, approximated by the surface area under the (engineering) stress–λ curve, where λ is the extension ratio, and λ_c denotes the extension ratio at the breaking point[30].



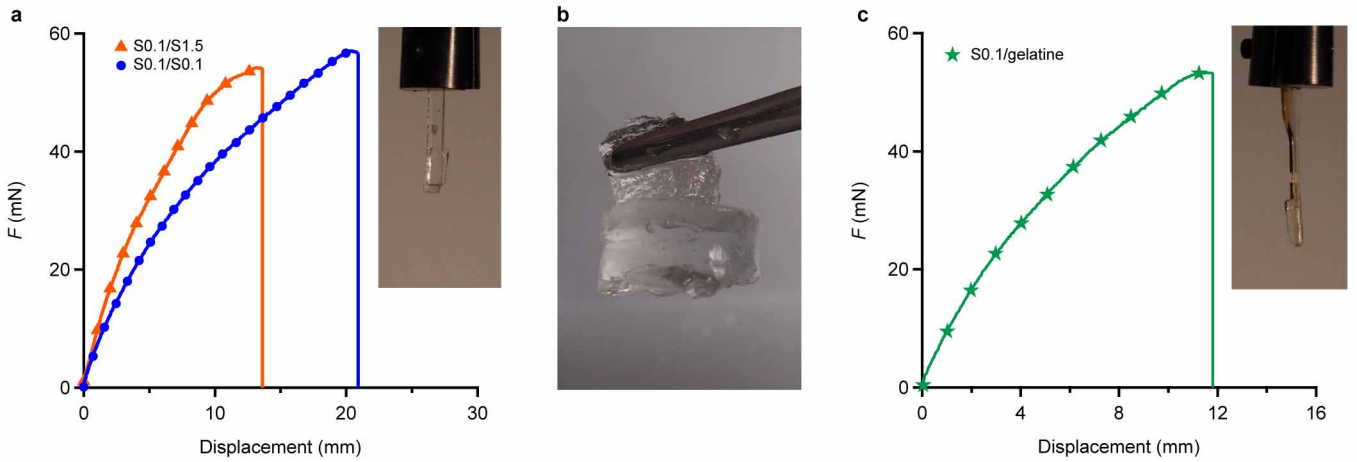
Extended Data Figure 1 | Lap-shear and tensile tests of PDMA S0.1 gels. **a**, Lap-joint geometry. Displacement was measured by a video extensometer that followed two markers (white dots), which were placed at a distance of 5 mm from the edge of the lap joint. The total length of the assembled ribbons was 40 mm. *w* denotes the width and *h* the thickness of gel ribbons. *l* is the overlap length. **b**, Comparison of force–displacement curves for PDMA S0.1 ribbon (black line) and for the lap joint glued by spreading 15 μl of TM-50 silica solution (blue circles). Displacement was measured by optical extensometer from two markers, initially spaced by 20 mm and centred on the joint. The PDMA S0.1/S0.1 assemblies broke outside the joint (inset). **c**, Lap-shear adhesion test reproducibility. Force–displacement curves of PDMA S0.1/S0.1 lap joint illustrating lap-shear test reproducibility. All gel ribbons were cut from the same plate. Bulk failure outside the joint was systematically observed.



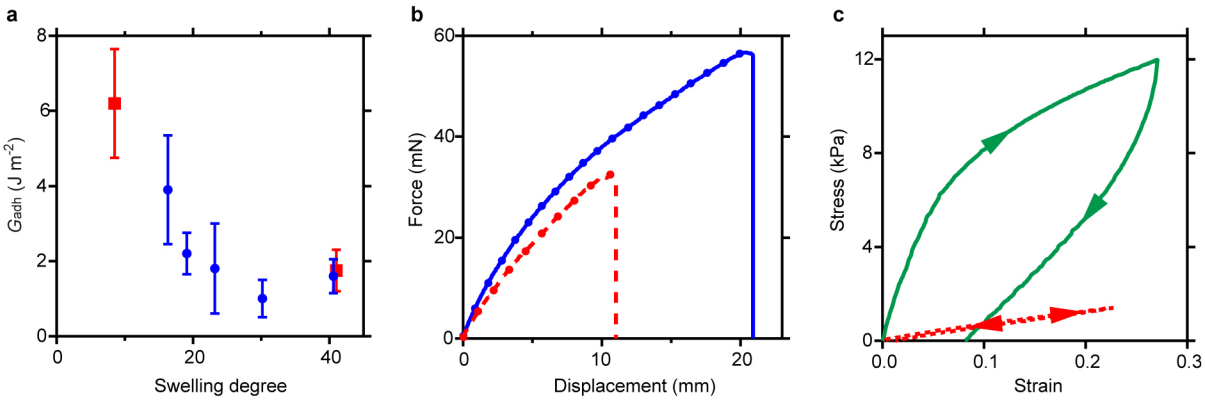
Extended Data Figure 2 | Measurement of adhesion energy of PDMA S0.1 gels. **a**, Lap-shear test geometry in which interfacial failure by peeling was observed for S0.1 gel ribbons glued by spreading 6 µl of TM-50 silica solution ($l = 5$ mm, $w = 2$ mm and $h = 5$ mm). **b**, Force–displacement curves for PDMA S0.1/S0.1 lap joints. Adhesive failure by interfacial peeling was observed. All gel ribbons were cut from the same gel plate and the tensile modulus was measured to be $E \approx 8.1 \pm 1.0$ kPa (error is s.d.). From the measured failure force, the adhesion energy can be estimated to be $6.2 \pm 1.4\ \mathrm{J\,m^{-2}}$ (error is s.d.).



Extended Data Figure 3 | Lap-shear test of PDMA S0.1 gels glued by various particle solutions. In order of increasing deformation at the breaking point are shown the adhesive failure by interfacial peeling in CNT-Thy (grey), CNC1 (green), SM-30 (yellow) and HS-40 (purple). Fracture outside the junction occurred for TM-50 (blue) and AL-30 (red). Lap-joint dimensions were $l = 5$ mm, $w = 5$ mm and $h = 2$ mm. 6 µl of solution was spread to make the junction.



Extended Data Figure 4 | Gluing gels of different stiffness or chemical nature. **a**, Force–displacement curve for an assembly made of soft PDMA S0.1 and rigid PDMA S1.5 (red triangles) gels glued by TM-50 solution. For comparison the results obtained under identical conditions for the symmetric PDMA S0.1/S0 assembly are plotted (blue circles). Lap-joint dimensions were $l = 10$ mm, $w = 5$ mm and $h = 2$ mm. 15 µl of TM-50 solution was spread to make the junction. **b**, Glued at their preparation state, both PDMA S0.1 and PDMA S1.5 gels initially had the same size (diameter of about 10 mm). The picture shows gels after 5 h of swelling in deionized water. The highly crosslinked PDMA S1.5 gel (top piece) is less swollen than the PDMA S0.1 gel (bottom piece). Interfacial stresses induced by heterogeneous overswelling considerably exceed the shear stresses applied in the mechanical lap-shear test of **a** and lead to interfacial failure during immersion and overswelling in water. De-bonding was slow. **c**, Lap-shear force–displacement trace for the gelatine and S0.1 PDMA gel assembly (green stars) glued by spreading TM-50 silica solution. The failure occurred outside the lap joint and the fatal crack propagated in tension mode.



Extended Data Figure 5 | Gluing overswollen gels and overswelling glued gels. **a**, Adhesion energy of joints made of PDMA S0.1 hydrogels swollen before being glued with AL-30 silica solutions to various degrees of swelling Q (in blue) and adhesion energy of joints made of S0.1 hydrogels glued with TM-50 silica solutions at as-synthesized swelling degree ($Q_0 \approx 8.5$) and after being immersed in water and swollen to reach the maximum, equilibrium swelling degree, $Q_e \approx 41$ (in red) (error bars are s.d.). When S0.1 gels were glued with AL-30 particles at the as-synthesized state, bulk failure systematically occurred outside the junction even when the joints were very short, narrow and thick. To induce peeling (interfacial failure) making cuts at the interface was necessary. **b**, Lap-shear test for PDMA S0.1/S0.1 assembly glued by TM-50 silica at the preparation state, Q_0 (blue circles) and after swelling in water for three days and attaining maximum equilibrium swelling Q_e (red circles). **c**, Mechanical behaviour of a nanocomposite sample that can be considered as a model of the interfacial adhesive layer. Engineering stress is plotted as a function of strain. The loading–unloading cycle at the gel preparation conditions (green line) shows a large hysteresis. At equilibrium swelling (red dashed line) the hysteresis and dissipation were very weak. Low dissipation at swelling equilibrium is responsible for a relatively weaker adhesion after immersion in water.

Extended Data Table 1 | Hydrogels

	Preparation state										Swelling equilibrium conditions		
Sample	Water (g)	DMA (g)	AAm (g)	MBA (mg)	KPS (mg)	TEMED (µl)	Silica v/v	Gelatine (g)	Q_0	E (kPa)	Q_e	E (kPa)	Swelling ratio Q_e/Q_0
S0.1	10.62	1.485		2.3	41	22.5			8.5	10 ± 1	41 ± 1	4 ± 1	5.1
NC	10.62	1.485		2.3	41	22.5	0.21		8.5	93 ± 8	29 ± 4	7 ± 1	3.4
S0.5	10.69	1.485		11.5	41	22.5			8.5	26 ± 1	23 ± 2		2.7
S1.0	10.77	1.485		23.0	41	22.5			8.5	45 ± 2	17 ± 2		2.0
S1.5	10.94	1.485		34.5	41	22.5			8.5	60 ± 4	15 ± 2		1.7
A0.1	7.62		1.065	2.3	41				9.0	11 ± 1			
Gelatine	11.10							3.31	5.3	30 ± 2	29 ± 2		4.9

Errors are s.d. AAm denotes acrylamide monomer, TEMED denotes *N,N,N',N'*-tetramethylethylenediamine.

LETTER

doi:10.1038/nature12781

Asymmetric synthesis from terminal alkenes by cascades of diboration and cross-coupling

Scott N. Mlynarski[1], Christopher H. Schuster[1] & James P. Morken[1]

Terminal, monosubstituted alkenes are ideal prospective starting materials for organic synthesis because they are manufactured on very large scales and can be functionalized via a broad range of chemical transformations. Alkenes also have the attractive feature of being stable in the presence of many acids, bases, oxidants and reductants. In spite of these attributes, relatively few catalytic enantioselective transformations have been developed that transform aliphatic α-olefins into chiral products with an enantiomeric excess greater then 90 per cent. With the exception of site-controlled isotactic polymerization of α-olefins[1], none of these catalytic enantioselective processes results in chain-extending carbon–carbon bond formation to the terminal carbon[2–6]. Here we describe a strategy that directly addresses this gap in synthetic methodology, and present a single-flask, catalytic enantioselective conversion of terminal alkenes into a number of chiral products. These reactions are facilitated by a neighbouring functional group that accelerates palladium-catalysed cross-coupling of 1,2-bis(boronates) relative to non-functionalized alkyl boronate analogues. In tandem with enantioselective diboration, this reactivity feature transforms alkene starting materials into a diverse array of chiral products. We note that the tandem diboration/cross-coupling reaction generally provides products in high yield and high selectivity (>95:5 enantiomer ratio), uses low loadings (1–2 mol per cent) of commercially available catalysts and reagents, offers an expansive substrate scope, and can address a broad range of alcohol and amine synthesis targets, many of which cannot be easily addressed with current technology.

Development of catalytic enantioselective reactions that operate efficiently with low catalyst loadings and high levels of selectivity is a paramount challenge in organic chemistry. This challenge is even greater when one targets the transformation of α-olefins that have a small steric bias between the prochiral faces of the alkene. For this reason, there are few catalytic asymmetric processes that operate effectively with aliphatic terminal alkenes. We sought to address this significant gap in synthesis methodology by developing a catalytic enantioselective reaction that converts terminal alkenes into chiral reactive intermediates; in this manner, one might introduce a number of useful catalytic asymmetric reactions simultaneously. A first step in the development of this strategy was achieved in engineering a Pt-catalysed enantioselective alkene diboration (Fig. 1a)[7]. In this Letter we present remarkably efficient cross-coupling reactions that apply to diboration products and collectively provide a strategy for enantioselective carbohydroxylation, carboamination and bisalkylation of terminal alkenes. These strategies enable the construction of many biologically significant molecules and should allow practicing chemists to assemble target structures in new ways. For example, the homoallylic alcohol embedded within the framework of the cytotoxic natural product epothilone C (Fig. 1b) might be accessed by a diboration/cross-coupling (DCC) reaction followed by oxidation. Alternatively, diboration followed by cross-coupling and amination could provide a new route to structural variants of the therapeutic agent tamsulosin from propene as a feedstock. Last, hydrocarbon stereocentres such as the one appearing in the antitumour macrolide kendomycin can be forged by DCC reaction followed by homologation of the remaining boronate.

The Pt-catalysed enantioselective diboration of terminal alkenes with $B_2(pin)_2$ (here pin indicates pinacolato) offers a platform for the construction of new molecular ensembles. In tandem with diboration, oxidation transforms terminal alkenes to enantiomerically-enriched 1,2-diols. A far greater range of new molecular building blocks would arise from terminal alkenes if 1,2-bis(pinacol boronates) would directly participate in efficient cross-coupling. Although related cross-couplings with bis(catechol boronates) are known[8], conversion of terminal alkenes to enantiomerically enriched 1,2-bis(catechol boronates) is generally not enantioselective. Therefore, a strategy for terminal alkene manipulation based on selective diboration reactions requires successfully engaging alkyl pinacol boronates as nucleophilic partners in Suzuki-Miyaura cross-coupling[9]. However, contrary to commonly employed

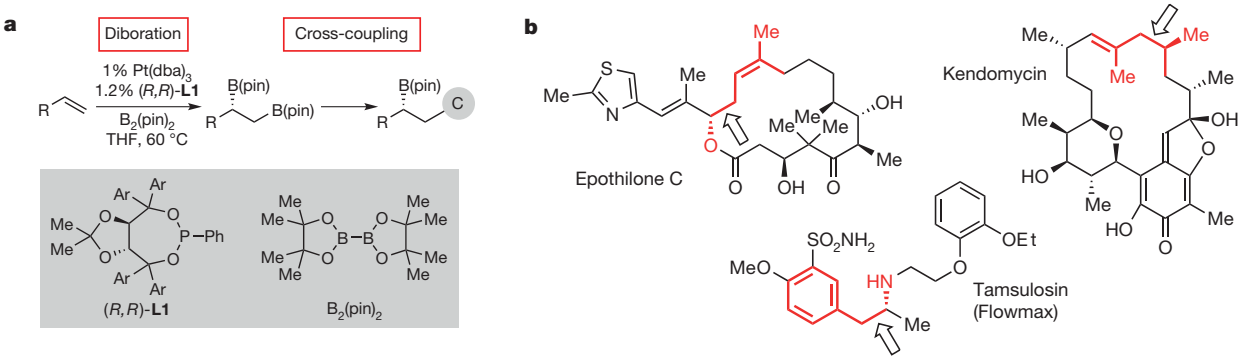


Figure 1 | The diboration/cross-coupling (DCC) strategy and potential applications. a, The DCC cascade. An efficient cross-coupling reaction that applies to alkyl pinacol boronates would enable conversion of terminal alkenes to a broad array of useful building blocks. **b**, Prospective targets. The DCC reaction followed by oxidation provides an alternative to carbonyl allylation for the construction of homoallylic alcohols as in epothilone C. Amination or homologation of the DCC product can provide access to chiral amines and simple chiral hydrocarbon building blocks. Open arrow indicates site of substrate alkene. Ar = 3,5-di-isopropylphenyl, dba = dibenzylideneacetone.

[1]Department of Chemistry, Merkert Chemistry Center, Boston College, Chestnut Hill, Massachusetts 02467, USA.

alkyl boranes and boronic acids, alkyl pinacol boronates are generally recalcitrant substrates in such processes[10]. Indeed, the only reported cross-coupling with a bis(pinacol boronate) involved two equivalents of a highly activated organic electrophile[11]. The contrasting reactivity between classes of boron reagents can be traced to a difference in transmetallation rates during the catalytic Suzuki cross-coupling reaction (Fig. 2a). Meticulous mechanistic studies[12,13] are in concert with prior assertions[9,14] and suggest that one operative mechanism for transmetallation involves pre-association of a Pd(hydroxide) with a neutral trivalent boron centre. Accordingly, it can be surmised that the diminished Lewis acidity of alkyl pinacol boronates relative to other boron derivatives lies at the root of the reactivity difference between these reagents. Thus, engaging pinacol alkyl boronates in cross-coupling has required the use of toxic bases[15] or pre-formed anionic four-coordinate "ate" complexes[16]. Recent advances[17] in the design of efficient ligands for metal-catalysed cross-coupling reactions have begun to provide a solution to this problem and indeed one recent report[18] suggests that pinacol boronates can participate in the Suzuki reaction in the presence of RuPhos[19], a monodentate phosphine ligand.

To determine whether 1,2-bis(pinacol boronates) can engage in efficient cross-couplings with unactivated organic electrophiles, we examined the reaction between isolated and purified bis(boronate) **1** (Fig. 2b) and bromobenzene under a wide range of reaction conditions. Optimal conditions are depicted in Fig. 2b and reveal that extraordinarily efficient cross-couplings can be achieved. In the presence of 1 mol% $Pd(OAc)_2$, 1 mol% RuPhos, and employing aqueous KOH as the base for 1 h, we obtained a 91% yield of alcohol **5** (Fig. 2b) after oxidative work-up. Importantly, alcohol **5** was isolated as a single constitutional isomer; the product from coupling with the secondary boronate was not detected. The reaction of **5** is remarkable in comparison to cross-couplings of other pinacol boronates (Fig. 2b). For example, under the same conditions in which **1** reaches 91% conversion, *n*-octyl boronate **2** is not detectably transformed. Reactions of 1,3-bis(boronate) **3** and 1,4-bis(boronate) **4** indicate that the rate acceleration experienced by **1** relies not just on the presence of a second boronate unit, but also on its position relative to the first.

To gain further insight into the special features of the 1,2-bis(pinacol boronate), we performed a direct competition experiment where both **1** and octylB(pin) **2** were subjected to cross-coupling in the same flask (Fig. 2c). In this experiment, >95% conversion of the 1,2-bisboronate was achieved whereas only a trace amount of product was produced from the monoboronate. With the reasonable assumption that transmetallation in the presence of hydroxide is irreversible, the outcome of this experiment suggests that the enhanced reactivity of the bis(boronate) is probably due to an enhanced rate of transmetallation; if transmetallation occurred at similar rates with each substrate but the rate retardation with octylB(pin) was due to slow reductive elimination, then octylB(pin) should sequester the Pd catalyst and retard the reaction of both substrates. Thus the presence of the adjacent non-reacting boronate appears to have a profound effect on the rate of transmetallation, leading to >50-fold enhancement in reactivity of the substrate.

A number of plausible explanations may account for the reaction acceleration by the vicinal boronate in **1**, but here we consider two. Similarly to the Lewis-base-induced activating effect that an adjacent carbonyl has on a reacting boronate[20], we considered that cooperative binding of hydroxide by the neighbouring boron centres might furnish an "ate" complex such as **A** (Fig. 3); for related observations in cross-coupling of geminal boronates, see ref. 21. Alternatively, we considered

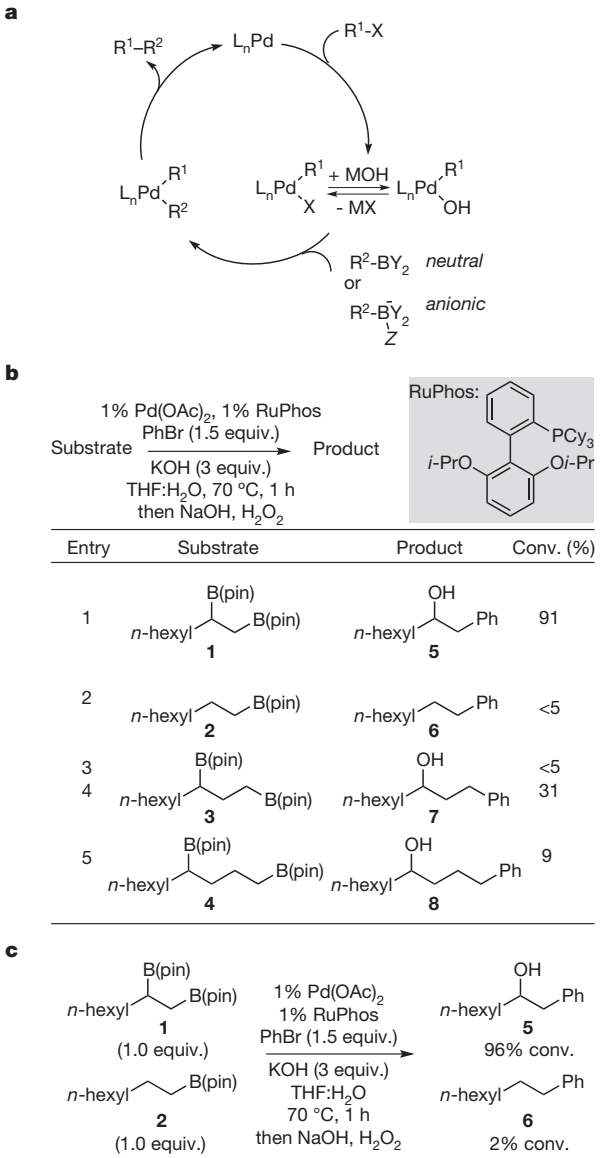


Figure 2 | Observations on the Pd-catalysed cross-coupling of 1,2-bis(boronates) with bromobenzene. **a**, Generalized mechanism for the palladium-catalysed Suzuki-Miyaura cross-coupling reaction. L_n = ligands. **b**, Cross-coupling of bromobenzene with alkyl pinacol boronates in the presence of RuPhos (structure shaded) shows a pronounced rate enhancement due to the presence of a vicinal boronate. Note that for entry 4, 5 mol% $Pd(OAc)_2$/RuPhos was employed for 20 h. Cy = cyclohexyl. **c**, A direct competition experiment suggests that the rate enhancement occurs at the first irreversible step or any step that precedes it.

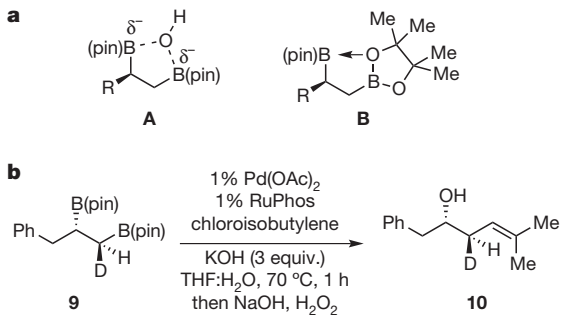


Figure 3 | Mechanistic considerations for the cross-coupling rate enhancement observed with 1,2-bis(boronates). **a**, Whereas rate enhancements in Suzuki-Miyaura couplings might result from internal Lewis base donation to the reacting boronate (as in compound **A**), internal Lewis acid activation (as in compound **B**) may allow the boronate to better bind a reactive Pd(OH) species. **b**, The stereochemical outcome of cross-coupling (retention of configuration at carbon) is consistent with an inner-sphere transmetallation, suggesting that internal Lewis acid activation is the more likely pathway.

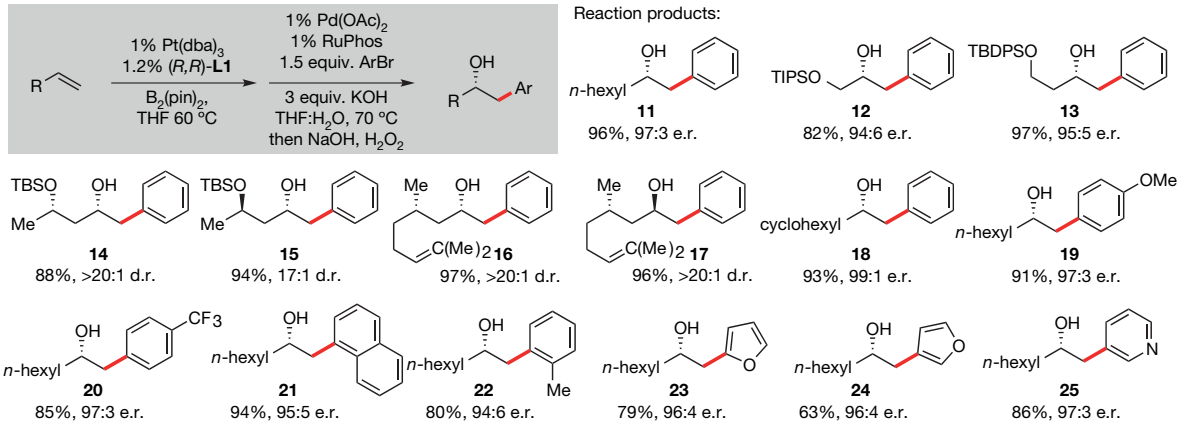


Figure 4 | Tandem single-pot DCC provides a new route to enantiomerically enriched benzylic alcohols from terminal alkenes. Shaded area shows the reaction: varying 'R' in the reactant gives reaction products shown. Under each product are given yield in per cent, and either enantiomeric ratio (e.r.) or diastereomer ratio (d.r.). Yield refers to isolated yield of purified product and is an average of two experiments (individual experimental values are within 10%). The e.r. was determined by chiral chromatography with an error $<\pm 2\%$. Note ligand (*S,S*)-**L1** was employed for compound **17**, and cross-coupling for **23** and **24** employed 1 equiv. LiCl. (*R,R*)-3,5-diethylphenyl-derived ligand was used in place of **L1** for **12**. $NaBO_3$ was used for oxidation of **12**, **14** and **15**. TIPS = triisopropylsilyl, TBDPS = *tert*-butyldiphenylsilyl, TBS = *tert*-butyldimethylsilyl.

that the function of the adjacent boron atom might be to act as a Lewis acid, coordinating to the pinacolato oxygen (**B**) thereby enhancing the Lewis acidity of the neighbouring boron centre; in line with the discussion above, this might enhance the reactivity of the primary organoboronate. The stereochemical outcome of the cross-coupling reaction might provide a clue to the operative reaction mechanism: with four-coordinate boron in **A**, transmetallation would necessarily occur by an outer-sphere path and result in inversion of configuration at the primary carbon[20]. Alternatively, reaction via **B** would most probably occur by an inner-sphere path and occur with retention of configuration[14,22]. To test these proposals, the cross-coupling of isotopically labelled substrate **9** and chloroisobutylene was examined and it was found to occur with retention of configuration at carbon, providing **10** as the product. The outcome of this labelling experiment suggests that an inner-sphere transmetallation may operate, perhaps involving association of a Pd(hydroxide) complex with **B**.

Conveniently, both alkene diboration and catalytic cross-coupling can be accomplished in a one-pot protocol, transforming simple terminal olefins to secondary alcohols after oxidation. Using bromobenzene as a model electrophile, 1-octene was found to successfully engage in the tandem sequence and provided homobenzylic alcohol **11** after oxidation of the cross-coupled product (Fig. 4). Replacement of bromobenzene with chlorobenzene also resulted in effective conversion to **11** (88% yield); however, reactions with either iodobenzene (30% yield) or phenyltriflate (48% yield) gave lower yields of the desired product. In addition to 1-octene, a number of other olefins were successfully engaged in the tandem sequence with bromobenzene as the electrophile, affording the products in high yield and with high enantioselectivity (products **12**–**18**). Both linear and branched aliphatic substrates as well as those containing pendant olefins and silyl ethers were well tolerated. Olefins derived from allylic or homoallylic alcohols underwent smooth reaction and importantly, when either hydrocarbon-based or oxygen-based β stereocentres are present (products **14**–**17**), effective catalyst control produces the products in excellent diastereomer ratios (17:1 to >20:1 d.r.). Examination of other aromatic electrophiles showed that electron-poor and electron-rich arenes as well as those that are sterically encumbered readily couple (**19**–**22**). Importantly, heteroaromatics can also be used and give highly enantioenriched adducts (**23**–**25**).

Homoallylic alcohols are strategically important compounds in synthetic organic chemistry, and significant resources have been directed towards their asymmetric construction. Almost exclusively, these motifs are accessed by allylation of carbonyl electrophiles with nucleophilic reagents[23]. An alternative route to these structures from terminal alkenes becomes available when vinyl electrophiles are engaged in the tandem DCC reaction. In preliminary studies, we investigated the coupling of vinyl bromides, but these most often occurred with lacklustre reaction efficiency (for example, 12% yield of **26** from the vinyl bromide). Whereas vinyl iodides were also ineffective (<5% yield), the reactions of vinyl chlorides were highly effective and furnished homoallylic alcohols in excellent yields on oxidative work-up (Fig. 5a). Because the olefin stereochemistry is retained during the course of Suzuki cross-coupling reactions, the DCC reaction provides ready access to homoallylic alcohols bearing configurationally defined trisubstituted (**27** and **28**), *cis*- and

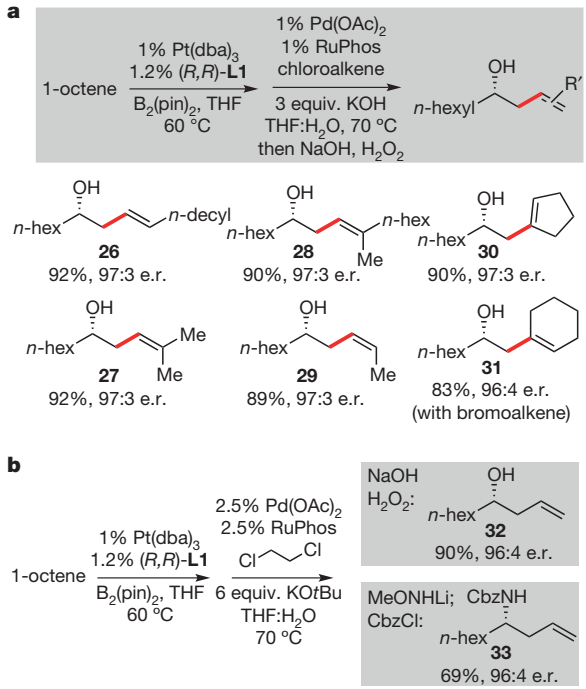


Figure 5 | The DCC tandem sequence provides access to synthetically useful chiral homoallylic alcohols that are not readily prepared by carbonyl allylation reactions. **a**, Construction of substituted homoallylic alcohols by DCC reaction/oxidation. Shaded area shows reaction: products are shown under. Variation of the chloroalkene employed in the cross-coupling gives differently substituted alkenes in the product. **b**, Use of dichloroethane allows for *in situ* formation of vinyl chloride and provides an effective route to unsubstituted homoallylic alcohols and amines (shaded). Yield and e.r. are as defined in Fig. 4 legend. Cbz = carboxybenzyl.

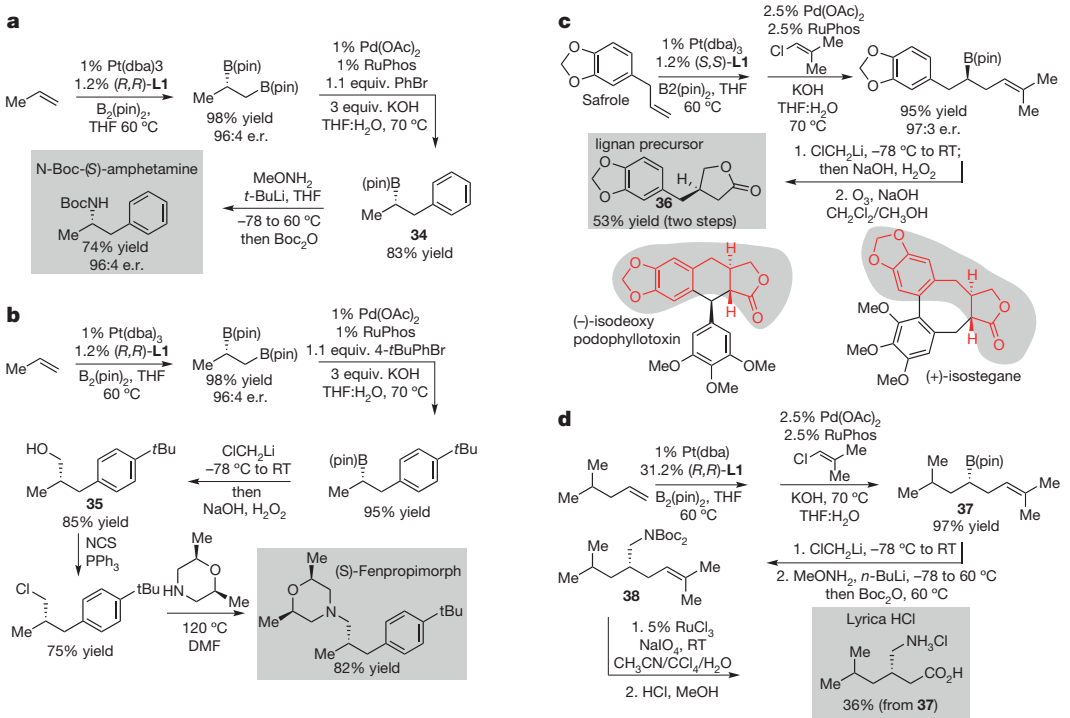


Figure 6 | DCC tandem reactions provide short new synthesis routes to important medicinal agents. These routes employ new feedstocks relative to existing routes and can facilitate new structure–activity relationship studies. **a**, Preparation of Boc protected (*S*)-amphetamine (shaded). **b**, Preparation of (*S*)-fenpropimorph. RT, room temperature. **c**, Construction of a key lignan building block **36**, which is a known precursor to isodeoxypodophyllotoxin and isostegane (red portion of the natural products shows the carbon framework of precursor **36**). **d**, Synthesis of non-racemic Lyrica.

trans-disubstituted (**26** and **29**) and cyclic alkenes (**30** and **31**) in a stereoselective fashion; construction of these substituted motifs with contemporary carbonyl allylation methods, when possible, requires specialized difficult-to-access reagents. Last, we considered that unsubstituted homoallylic alcohols might be accessed from vinyl chloride; however, handling this toxic and gaseous electrophile is cumbersome and requires specialized equipment. We found that a straightforward alternative arises from the use of dichloroethane: under the basic reaction conditions, this inexpensive liquid reagent is presumably converted to vinyl chloride and engages in cross-coupling (Fig. 5b). After oxidative work-up, we isolated unsubstituted homoallylic alcohol **32** in outstanding yield and enantiomeric excess. Of note, homoallylic amines can also be accessed with the DCC strategy by subjecting the purified DCC intermediate to amination rather than *in situ* oxygenation. This allowed construction of **33** from 1-octene (Fig. 5b) also in outstanding levels of enantioselectivity and good yield; for recent efficient catalytic enantioselective imine allylation to give allyl amines see ref. 24.

The versatility of the catalytic DCC reaction provides a rapid route to important product motifs from simple alkene feedstocks. To demonstrate the power of this strategy, we targeted a diverse array of biologically important molecules. For example, chiral phenethylamine derivatives are broadly active pharmaceutical agents most often produced from ketone precursors[25]. Alternatively, as the sequence in Fig. 6a indicates, the DCC reaction allows N-*tert*-butoxycarbonyl-(*S*)-amphetamine to be prepared quickly from the commodity chemical propene as the starting material. Efficient diboration is followed by cross-coupling to give secondary boronate **34**. Subjection of **34** to stereospecific direct amination[26] and Boc protection furnished the target in excellent enantiomeric purity. Importantly, the modular nature of this new route to phenethylamine derivatives could dovetail with strategies for high-throughput synthesis and provide rapid access to new derivatives for biological studies. Using the one-pot DCC in conjunction with boronate homologation[27] and oxidation converts (Fig. 6b) propene to primary alcohol **35**. Chlorination followed by S_N2 reaction furnished the potent fungicide (*S*)-fenpropimorph with high levels of enantioenrichment (Fig. 6b)[28]. Another application involves construction of lignan lactones, a broad class of natural products that exhibit a wide range of biological activity. Starting from commercially available safrole, we constructed key intermediate **36** in three steps: DCC reaction, boronate homologation and oxidative olefin cleavage (Fig. 6c). Simple alkylation of lactone **36** is known to furnish a variety of lignan natural products such as isodeoxypodophyllotoxin and isostegane[29]. Last, we note that the DCC reaction can be used to give **37**; homologation, amination and protection of **37** furnished **38** (Fig. 6d). Subsequent oxidative olefin cleavage and deprotection of **38** provides a new route to the pharmaceutical agent Lyrica (pregabalin) from a unique set of organic chemical building blocks (Fig. 6d)[30].

In summary, the catalytic enantioselective diboration of terminal alkenes, combined with Pd-catalysed cross-coupling, provides a flexible platform for the construction of a broad array of chiral compounds from non-functionalized terminal alkenes. Although application of this methodology in target-oriented synthesis is easy to foresee, when one considers the tremendous variety of α-olefins and aryl/vinyl electrophiles that are available, the DCC reaction sequence should also provide new strategies for diversity-based synthesis. In addition to its direct impact on molecular synthesis, the studies presented here define a unique reactivity characteristic of 1,2-bis(pinacol boronates). We anticipate that the enhanced reactivity of the 1,2-bis(boronate) in transmetallation reactions may enable a range of other important enantioselective terminal alkene transformations that will have an impact on the field of chemical synthesis.

METHODS SUMMARY

The general procedure for the one-pot DCC reaction as described in Fig. 4 is as follows. $Pt(dba)_3$ (1.0 mol%), (*R,R*)-**L1** (1.2 mol%), $B_2(pin)_2$ (1.05 equiv.) and anhydrous THF ([substrate] = 1.0 M) are stirred together at 80 °C for 15 min. After cooling to ambient temperature, the alkene (1.0 equiv.) is added and the reaction mixture is stirred at 60 °C for 3 h. On cooling to ambient temperature, $Pd(OAc)_2$ (1.0 mol%), followed by RuPhos (1.0 mol%), the electrophile (1.5 equiv.), KOH (3.0 equiv.), additional THF and deoxygenated water ([substrate] = 0.1 M; 10:1 v:v

THF:H_2O) are added and the reaction mixture is heated to 70 °C for 12 h. The reaction is then cooled to 0 °C and treated with 3 M aqueous NaOH and 30% H_2O_2. After 4 h at ambient temperature, excess H_2O_2 is carefully quenched with saturated aqueous $Na_2S_2O_3$, followed by extraction with ethyl acetate. The combined organics are dried over Na_2SO_4, filtered and concentrated. The resulting material is purified by flash chromatography on silica gel. For complete experimental details and characterization of all new compounds, see Supplementary Information.

Received 7 August; accepted 15 October.
Published online 18 December 2013.

1. Resconi, L., Cavallo, L., Fait, A. & Piemontesi, F. Selectivity in propene polymerization with metallocene catalysts. *Chem. Rev.* **100,** 1253–1345 (2000).
2. Subbarayan, V., Ruppel, J. V., Zhu, S., Perman, J. A. & Zhang, X. P. Highly asymmetric cobalt-catalyzed aziridination of alkenes with trichloroethoxysulfonyl azide ($TcesN_3$). *Chem. Commun.* 4266–4268 (2009).
3. Uozumi, Y. & Hayashi, T. Catalytic asymmetric synthesis of optically active 2-alkanols via hydrosilylation of 1-alkenes with a chiral monophosphine-palladium catalyst. *J. Am. Chem. Soc.* **113,** 9887–9888 (1991).
4. Lo, M. M.-C. & Fu, G. C. A new class of planar–chiral ligands: synthesis of a C_2-symmetric bisazaferrocene and its application in the enantioselective Cu(I)-catalyzed cyclopropanation of olefins. *J. Am. Chem. Soc.* **120,** 10270–10271 (1998).
5. Becker, H. & Sharpless, K. B. A new ligand class for the asymmetric dihydroxylation of olefins. *Angew. Chem. Int. Edn Engl.* **35,** 448–451 (1996).
6. Noonan, G. M., Fuentes, J. A., Cobley, C. J. & Clarke, M. L. An asymmetric hydroformylation catalyst that delivers branched aldehydes from alkyl alkenes. *Angew. Chem. Int. Edn Engl.* **51,** 2477–2480 (2012).
7. Coombs, J. R., Haefner, F., Kliman, L. T. & Morken, J. P. Scope and mechanism of the Pt-catalyzed enantioselective diboration of monosubstituted alkenes. *J. Am. Chem. Soc.* **135,** 11222–11231 (2013).
8. Miller, S. P., Morgan, J. B., Nepveux, F. J. & Morken, J. P. Catalytic asymmetric carbohydroxylation of alkenes by a tandem diboration/Suzuki cross-coupling/oxidation reaction. *Org. Lett.* **6,** 131–133 (2004).
9. Miyaura, N. & Suzuki, A. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.* **95,** 2457–2483 (1995).
10. Jana, R., Pathak, T. P. & Sigman, M. S. Advances in transition metal (Pd, Ni, Fe)-catalyzed cross coupling reactions using alkyl-organometallics as reaction partners. *Chem. Rev.* **111,** 1417–1492 (2011).
11. Lee, Y., Jang, H. & Hoveyda, A. H. Vicinal diboronates in high enantiomeric purity through tandem site-selective NHC-Cu-catalyzed boron-copper additions to terminal alkynes. *J. Am. Chem. Soc.* **131,** 18234–18235 (2009).
12. Carrow, B. P. & Hartwig, J. F. Distinguishing between pathways for transmetallation in Suzuki-Miyaura reactions. *J. Am. Chem. Soc.* **133,** 2116–2119 (2011).
13. Amatore, C., Jutand, A. & Le Duc, G. Kinetic data for the transmetalation/reductive elimination in palladium-catalyzed Suzuki–Miyaura reactions: unexpected triple role of hydroxide ions used as base. *Chemistry* **17,** 2492–2503 (2011).
14. Matos, K. & Soderquist, J. A. Alkylboranes in the Suzuki-Miyaura coupling: stereochemical and mechanistic studies. *J. Org. Chem.* **63,** 461–470 (1998).
15. Sato, M., Miyaura, N. & Suzuki, A. Cross-coupling reaction of alkyl- or arylboronic acid esters with organic halides induced by thallium(I) salts and palladium catalyst. *Chem. Lett.* **18,** 1405–1408 (1989).
16. Zou, G. & Falck, J. R. Suzuki-Miyaura cross coupling of lithium *n*-alkylborates. *Tetrahedr. Lett.* **42,** 5817–5819 (2001).
17. Christmann, U. & Vilar, R. Monoligated palladium species as catalysts in cross-coupling reactions. *Angew. Chem. Int. Edn Engl.* **44,** 366–374 (2005).
18. Yang, C.-T. *et al.* Alkylboronic esters from copper-catalyzed borylation of primary and secondary alkyl halides and pseudohalides. *Angew. Chem. Int. Edn Engl.* **51,** 528–532 (2012).
19. Charles, M. D., Schultz, P. & Buchwald, S. L. Efficient Pd-catalyzed amination of heteroaryl halides. *Org. Lett.* **7,** 3965–3968 (2005).
20. Sandrock, D. L., Jean-Gérard, L., Chen, C., Dreher, S. D. & Molander, G. A. Stereospecific cross-coupling of secondary alkyl β-trifluoroboratoamides. *J. Am. Chem. Soc.* **132,** 17108–17110 (2010).
21. Endo, K., Ohkubo, T., Hirokami, M. & Shibata, T. Chemoselective and regiospecific Suzuki coupling on a multisubstituted sp^3 carbon in 1,1-diborylalkanes at room temperature. *J. Am. Chem. Soc.* **132,** 11033–11035 (2010).
22. Ridgway, B. H. & Woerpel, K. A. Transmetalation of alkylboranes to palladium in the Suzuki coupling reaction proceeds with retention of stereochemistry. *J. Org. Chem.* **63,** 458–460 (1998).
23. Yus, M., González-Gómez, J. C. & Foubelo, F. Catalytic enantioselective allylation of carbonyl compounds and imines. *Chem. Rev.* **111,** 7774–7854 (2011).
24. Silverio, D. L. *et al.* Simple organic molecules as catalysts for enantioselective synthesis of amines and alcohols. *Nature* **494,** 216–221 (2013).
25. Nogrady, T. & Weaver, D. F. (eds) *Medicinal Chemistry: A Molecular and Biochemistry Approach* Ch. 4 193–309 (Oxford, 2005).
26. Mlynarski, S. N., Karns, A. S. & Morken, J. P. Direct stereospecific amination of alkyl and aryl pinacol boronates. *J. Am. Chem. Soc.* **134,** 16449–16451 (2012).
27. Sadhu, K. M. & Matteson, D. S. (Chloromethyl)lithium: efficient generation and capture by boronic esters and a simple preparation of diisopropyl (chloromethyl)boronate. *Organometallics* **4,** 1687–1689 (1985).
28. Himmele, W. & Pommer, E.-H. 3-Phenylpropylamines: a new class of systemic fungicides. *Angew. Chem. Int. Edn Engl.* **19,** 184–189 (1980).
29. Itoh, T., Chika, J., Takagi, Y. & Nishiyama, S. An efficient enantioselective total synthesis of antitumor lignans: synthesis of enantiomerically pure 4-hydroxyalkanenitriles via an enzymatic reaction. *J. Org. Chem.* **58,** 5717–5723 (1993).
30. Silverman, R. B. & Andruszkiewicz, R. Gamma amino butyric acid analogs and optical isomers. US Patent 6,197,819 B1 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by the US National Institutes of Health, Institute of General Medical Sciences (grant GM-59417). S.N.M. and C.H.S. were supported by John LaMattina graduate fellowships. We thank Allychem for providing $B_2(pin)_2$.

Author Contributions S.N.M. and C.H.S. developed the procedure for the DCC reaction and collected the data in Figs 2b and 4. S.N.M. conducted the studies in Figs 5 and 6 and conducted the isotope labelling experiment in Fig. 3. J.P.M. conceived and designed the studies, planned the research and wrote the manuscript with assistance from C.H.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interest. Correspondence and requests for materials should be addressed to J.P.M. (morken@bc.edu).

LETTER

doi:10.1038/nature12913

A signature of transience in bedrock river incision rates over timescales of 10^4–10^7 years

Noah J. Finnegan[1], Rina Schumer[2] & Seth Finnegan[3]

Measured rates of river incision into bedrock are commonly interpreted as proxies for rates of rock uplift (see refs 1 and 2, for example) and indices of the strength of climatic forcing of erosion over time (see refs 3 and 4, for example). This approach implicitly assumes that river incision rates are in equilibrium with external forcings over a wide range of timescales. Here we directly test this assumption by examining the temporal scaling of bedrock river incision from 155 independent measurements of river incision compiled from 14 sites. Of these sites, 11 exhibit a negative power-law dependence of bedrock river incision rate on measurement interval, a relationship that is apparent over timescales of 10^4–10^7 years and is independent of tectonic and geomorphic setting. Thus, like rates of sediment accumulation[5], rates of river incision into bedrock exhibit non-steady-state behaviour even over very long measurement intervals. Non-steady-state behaviour can be explained by episodic hiatuses in river incision triggered by alluvial deposition, if such hiatuses have a heavy-tailed length distribution[6]. Regardless of its cause, the dependence of incision rate on measurement interval complicates efforts to infer tectonic or climatic forcing from changes in rates of river incision over time or from comparison of rates computed over different timescales.

Unglaciated topography is shaped by competition between tectonic uplift and bedrock river incision[7]. The potential for rivers to grow steeper, convey more water via orographic precipitation and thus become more erosive with higher rates of tectonic uplift suggests that rates of surface erosion should evolve to match rates of rock uplift in actively uplifting ranges[8]. At the same time, climate-driven changes in sediment supply and water discharge to rivers are thought to modulate rates of vertical river incision into rock over geologic time[9,10]. These arguments imply that measured rates of bedrock river incision can constrain active tectonic processes as well as temporal variability in the strength of climate forcing of erosion.

Measured rates of land-surface change in aggradational (formed by sediment deposits) settings, in contrast, exhibit a negative power-law dependence of accumulation rate on measurement interval that is not directly attributable to tectonic or climatic forcing[11]. This 'Sadler effect' arises when the duration–frequency distribution of hiatuses between intervals of accumulation has a heavy-tailed distribution[6]. Such distributions emerge from a variety of stochastic sediment accumulation models[6]. Over measurement intervals smaller than the longest hiatus, sedimentary sequences incorporate longer hiatuses at longer timescales and thus average accumulation rates tend to decline with measurement interval. It is unclear whether information about external climatic or tectonic forcing can then be recorded[6]. For example, an analysis of clastic coastal and continental shelf deposits indicates that only for measurement intervals exceeding 10^4–10^5 years do rates of sediment accumulation cease declining and therefore begin to reflect tectonic subsidence rates[5]. This measurement interval corresponds to the longest recorded hiatus.

A bedrock river is commonly conceptualized as an alluvial river bed overlying a bedrock channel bed that is incised only when the stripping of deposited alluvial material exposes bedrock to processes of abrasion, weathering and plucking[12]. Hiatuses in incision begin when the alluvial bed aggrades and end when its elevation returns to that of the bedrock channel, for example following flood scour[12] (Fig. 1a, b). If the long time series of alluvial bed elevation can be described as a stochastic process[6,12], then the return time to bedrock will follow a power-law frequency distribution in the long time limit[6]. This is true whether aggradation and scour events follow a simple random walk process[13], a long-range correlated random process, or a random walk marked by power-law periods between deposition or erosional events[14]. Under this model, bedrock incision rate will decline with measurement interval according to a negative power-law relationship (Fig. 1c). The measurement interval over which a system exhibits negative power-law rate scaling is related to the longest physically possible hiatus duration in the period of record. Only over timescales exceeding the duration of the longest possible hiatus can changes in river incision rate be confidently interpreted in terms of tectonic or climatic forcing.

It is widely recognized that mass-wasting triggers rapid and deep alluvial bed aggradation in bedrock channels following earthquakes[15], fires[16] and large storms[17]. Additionally, both rapid scour and deposition of alluvium occurs during floods in bedrock channels. Given the stochastic nature of these processes governing alluvial bed elevation change[18], the potential exists for a negative power-law dependence of bedrock river incision rates on measurement interval. Although numerical modelling suggests that steady-state bedrock incision arises over relatively short measurement intervals (10^3 years) despite stochastic forcing of sediment supply[12], a review of bedrock incision records in the southeastern USA[19] observed a negative scaling of bedrock incision rates with measurement intervals of 10^3–10^7 years. Another study observed negative rate scaling for a variety of erosion processes over similar timescales[20]. Therefore, the fidelity of the process of bedrock river incision as a recorder of external climatic and tectonic forcing remains uncertain.

To test for a dependency of bedrock river incision rate on measurement interval, we compiled 14 bedrock river incision data sets that each span at least one order of magnitude in time (see Methods and Supplementary Table 1). To avoid the spurious correlation that arises from plotting a rate against its own denominator (measurement interval), we computed the power-law relationship between cumulative bedrock incision and measurement interval for each of the 14 data sets (see Methods). The temporal scaling of cumulative bedrock incision can then be related to the temporal scaling of bedrock incision rate by subtracting one from the cumulative incision versus measurement interval power-law exponent, herein referred to as β (Fig. 1c)[5]. Of the 14 data sets, 11 exhibit values of β that are less than one, implying a negative power-law dependence of incision rate on measurement interval (Fig. 2; Extended Data Table 1). For the entire data set, the mean power-law exponent relating cumulative river incision and measurement interval is about 0.8 (implying a rate versus measurement interval exponent of about −0.2). In addition, we find that the apparent negative power-law dependence of incision rate on measurement interval persists over four orders of magnitude in time (10^4–10^7 years) (Fig. 3). Because tectonically inactive rivers tend to preserve longer incision records than tectonically active

[1]Department of Earth and Planetary Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA. [2]Division of Hydrologic Sciences, Desert Research Institute, Reno, Nevada 89512, USA. [3]Department of Integrative Biology, University of California Berkeley, Berkeley, California 94720, USA.

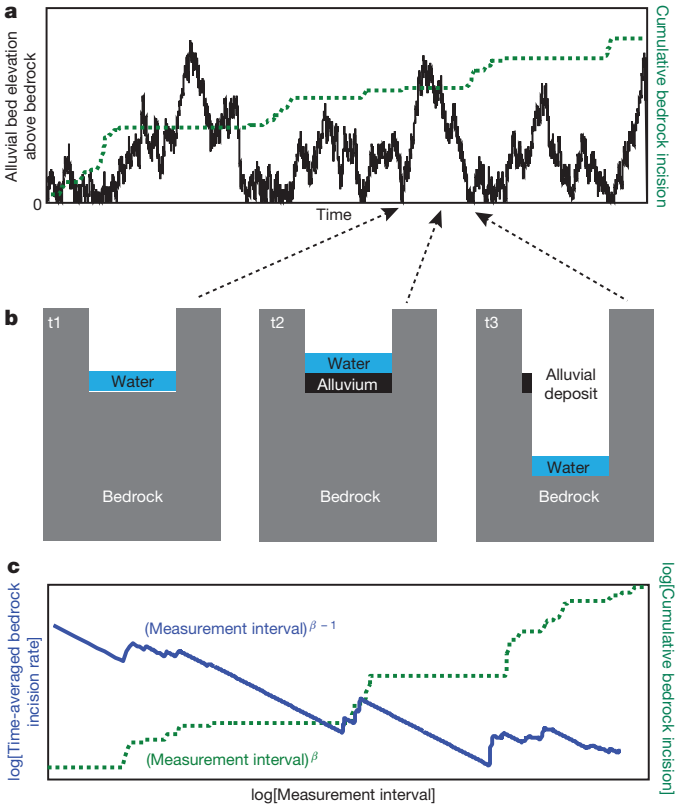


Figure 1 | The connection between stochastic alluvial bed elevation change, cumulative bedrock incision and incision rate scaling. **a**, Random-walk simulation of alluvial bed elevation (black line) and cumulative bedrock incision (dashed green line), assuming incision rate is constant when the bed is unburied. **b**, Bedrock canyon evolution implied by simulation. **c**, Plot of the logarithm of cumulative incision (dashed green line) and the logarithm of time-averaged incision rate (solid blue line) versus the logarithm of the measurement interval calculated from the synthetic incision record in **a**. Adjacent to each curve is the form of the power-law relationship of that variable (cumulative bedrock incision or time-averaged bedrock incision rate) to measurement interval in terms of the power-law exponent relating cumulative incision to measurement interval, β.

rivers, it is impossible to completely de-convolve measurement interval and tectonic setting. Nevertheless, we see no evidence to suggest that the observed scaling is strongly influenced by tectonic setting (Fig. 3; Extended Data Table 1). Lastly, we note that a negative power-law dependence of incision rate on measurement interval is apparent regardless of the particular landform used to constrain river incision (Extended Data Table 1). Because preservation of paired strath terraces is promoted by channel narrowing, which is a common response to accelerating incision[21], it is conceivable that the terrace record could be biased towards settings with accelerating incision and hence negative power-law rate scaling. However, a negative power-law dependence of incision rate on measurement interval is also recorded in unpaired terraces, caves and incised volcanic deposits, which should not be subject to the same preservation bias because they do not require valley narrowing for preservation. Consequently, it is unlikely that our findings simply reflect a preservation bias.

Rates of bedrock incision recorded during floods[22] and over short measurement intervals[23] are usually too large to be sustained over geologic timescales, implying that long hiatuses must separate intervals of incision. Evidence that such hiatuses have a power-law distribution in time (and are therefore probably stochastic in origin) is supported by the negative power-law dependence of incision rate on measurement interval that we observe over 10^4–10^7 years. Physical evidence for extremely long incision hiatuses in bedrock exists in the Himalayas, where valley fills can persist for over 10^5 years (ref. 24). Therefore it is possible that stochastic forcing from processes of sediment transport and delivery may be recorded in

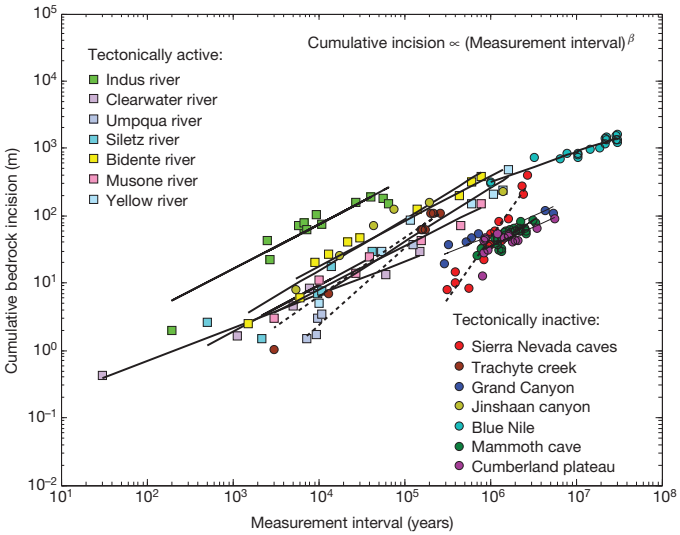


Figure 2 | Cumulative bedrock incision as a function of measurement interval. Log–log plot of cumulative incision versus measurement interval for the 14 data sets. Best-fit lines reflect the mean slopes and intercepts derived from the error analysis. Dashed lines indicate slopes (that is, β) greater than or equal to one. Solid lines indicate slopes less than one.

rates of bedrock incision even over measurement intervals exceeding 10^5 years.

It has also been argued that gravel aggradation modulated by Quaternary climate changes provides an external mechanism for generating incisional hiatuses in bedrock channels[4,10]. In this case, return time distribution theory would predict a levelling out of incision rates at the period of the forcing and thus a transition to an exponent of 1 on a plot of cumulative incision versus measurement interval. For example, in numerical experiments in which incision hiatuses are driven by glacial–interglacial climate cycles, rates of vertical incision show little change once averaged over the 10^5-year interval that corresponds to the dominant period of late Pleistocene glaciations[10]. In contrast, we find that the negative power-law dependence of incision rate on measurement interval persists over four orders of magnitude in time (10^3–10^7 years) (Fig. 3), and—importantly—over intervals much longer than the period of any known periodic climate forcing. This suggests that, globally, hiatuses in river incision into bedrock may not be coupled in a simple linear way to periodic Pleistocene climate forcing, as has been frequently suggested.

If incisional hiatus length is instead imposed by something intrinsic to the process that generates hiatuses (for example, maximum landslide

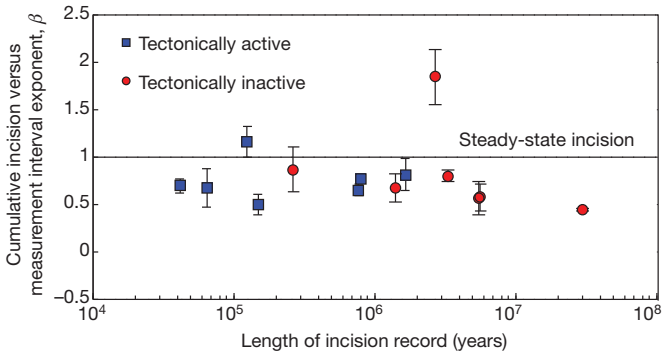


Figure 3 | Cumulative bedrock incision versus measurement interval exponent β as a function of incision record length. Data points are the mean exponent calculated from the error analysis. Error bars represent $\pm 1\sigma$ derived from the error analysis. Data points plotted above the steady-state incision line reflect rates of incision that increase with measurement interval. Data points plotted below the steady-state incision line reflect rates of incision that decrease with measurement interval.

size), there is no evidence to support the existence of such a limit in the data examined here. Long-term average rates of bedrock incision are thus apparently inseparable from the interval over which they are measured. Before incision rates at different locations can be meaningfully compared, the rate-measurement interval scaling of each location must first be quantified and then the incision rates need to be normalized to a particular measurement interval. The data compiled here also suggest that comparisons between river incision rates and, for example, thermochronologic ages or geodetic rates of surface uplift may be complicated by the fact that incision rates depend on measurement interval, as well as tectonics and climate, as also argued by ref. 20. However, to the extent that stochastic variation in sediment input to channels is regionally coherent[25], the incision histories of nearby catchments might be quite similar, even if they do not record tectonic or climatic forcing in a straightforward way.

Our inference that the dependence of incision rate on measurement interval reflects the influence of episodic alluvial aggradation is supported by the observation that bedrock river channels are usually covered in debris, even in tectonically active settings[12,26,27]. The incision scaling observed on the Blue Nile river[28], if it arises from power-law hiatuses, implies an instantaneous incision of about 0.7 m in 1 year, but only 116 m after 10^5 years. Thus, only during 0.2% of a 10^5-year-long record (that is, 166 years) would any channel incision occur. In other words, as observers, we are far more likely to encounter a channel in a state of non-incision. This suggests that an understanding of the processes that generate incisional hiatuses is arguably more important to understanding rates of landscape change than is an understanding of the incision process itself, as has also been argued by ref. 12. Our analysis also suggests that apparent recent increases in river incision rates should be an expected consequence of measuring bedrock river incision rates over measurement intervals that do not incorporate long hiatuses, as argued by ref. 10.

Do the same issues that complicate the interpretation of river incision rates also affect landscape-scale erosion rates? A study of sediment yields in steep, mountainous catchments on decadal measurement intervals suggests that landscape-scale erosion rates (as opposed to fluvial incision rates) are biased towards slow rates over short measurement intervals because of infrequently sampled catastrophic erosion events[25]. This scaling is the opposite of what we observe in the case of river incision, where shorter measurement intervals tend to yield higher rate estimates. One potential explanation for this apparent contradiction is that materials eroded from hill slopes are subsequently deposited in channels, where they may cause hiatuses in bedrock channel incision. Infrequent, large hill-slope erosion events could therefore trigger infrequent but long-duration hiatuses in bedrock channel incision. That said, in general landscape-scale erosion rates show little apparent dependence on measurement interval[29], an observation that has been attributed to the averaging out of local stochasticity with increasing spatial scale of measurement[29]. Thus, the influence of locally stochastic erosional processes on temporal rate scaling may depend both on the specific process in question and on the spatial scale examined.

METHODS SUMMARY

We compiled 14 bedrock river incision data sets that span at least one order of magnitude in time (Supplementary Table 1), encompassing a total of 155 measurements of river incision. We define tectonically active settings as those regions with documented tectonically driven rock uplift. Each of the tectonically active data sets was also taken from a paper (see the Extended Methods for citations) that measured river incision to constrain active tectonic processes. Tectonically inactive settings did not meet these criteria. We use reported or estimated uncertainties for each measurement of cumulative bedrock incision and measurement interval to define the uncertainty for each data point in each data set. We performed a Monte Carlo error analysis for each data set in which we calculated 3,000 linear fits between the logarithm of cumulative bedrock incision and the logarithm of measurement interval. Measurement interval for a cumulative river incision measurement is equivalent to the age of the landform used to constrain incision. This is because cumulative river incision for all data points is computed between the modern channel elevation and the palaeo-channel elevation. We do not calculate incision between dated landforms because such estimates do not represent statistically independent measurements of incision. For each model iteration, we assigned incision and age errors by selecting randomly from a normal error distribution for each data point with a standard deviation corresponding to the reported or estimated uncertainty in landform age and elevation. Slope and intercept distributions were created for each data set from the results of the Monte Carlo simulation to define power-law exponents relating cumulative incision and measurement interval, as well as corresponding uncertainties. Because our explanatory variable (measurement interval) has significant uncertainty, we use a total-least-squares regression method to quantify slope.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 June; accepted 25 November 2013.

1. Pazzaglia, F. J. & Brandon, M. T. A fluvial record of long-term steady-state uplift and erosion across the Cascadia forearc high, western Washington State. *Am. J. Sci.* **301,** 385–431 (2001).
2. Burbank, D. W. *et al.* Bedrock incision, rock uplift and threshold hillslopes in the northwestern Himalayas. *Nature* **379,** 505–510 (1996).
3. Leland, J., Reid, M. R., Burbank, D. W., Finkel, R. & Caffee, M. Incision and differential bedrock uplift along the Indus River near Nanga Parbat, Pakistan Himalaya, from ^{10}Be and ^{26}Al exposure age dating of bedrock straths. *Earth Planet. Sci. Lett.* **154,** 93–107 (1998).
4. Bull, W. L. & Knuepfer, P. L. K. Adjustments by the Charwell River, New Zealand, to uplift and climatic changes. *Geomorphology* **1,** 15–32 (1987).
5. Jerolmack, D. J. & Sadler, P. Transience and persistence in the depositional record of continental margins. *J. Geophys. Res.* **112,** F03S13 (2007).
6. Schumer, R. & Jerolmack, D. J. Real and apparent changes in sediment deposition rates through time. *J. Geophys. Res.* **114,** F00A06 (2009).
7. Howard, A. D., Dietrich, W. E. & Seidl, M. A. Modeling fluvial erosion on regional to continental scales. *J. Geophys. Res.* **99,** 13971–13986 (1994).
8. Whipple, K. X. Bedrock rivers and the geomorphology of active orogens. *Annu. Rev. Earth Planet. Sci.* **32,** 151–185 (2004).
9. Molnar, P. *et al.* Quaternary climate change and the formation of river terraces across growing anticlines on the north flank of the Tien Shan, China. *J. Geol.* **102,** 583–602 (1994).
10. Hancock, G. S. & Anderson, R. S. Numerical modeling of fluvial strath-terrace formation in response to oscillating climate. *Geol. Soc. Am. Bull.* **114,** 1131–1142 (2002).
11. Sadler, P. M. Sediment accumulation rates and the completeness of stratigraphic sections. *J. Geol.* **89,** 569–584 (1981).
12. Lague, D. Reduction of long-term bedrock incision efficiency by short-term alluvial cover intermittency. *J. Geophys. Res.* **115,** F02011 (2010).
13. Redner, S. *A Guide to First-Passage Processes* Chs 1 and 3 (Cambridge Univ. Press, 2007).
14. Ding, M. Z. & Yang, W. M. Distribution of the first return time in fractional Brownian-motion and its application to the study of on-off intermittency. *Phys. Rev. E* **52,** 207–213 (1995).
15. Yanites, B. J., Tucker, G. E., Mueller, K. J. & Chen, Y.-G. How rivers react to large earthquakes: evidence from central Taiwan. *Geology* **38,** 639–642 (2010).
16. Roering, J. J. & Gerber, M. Fire and the evolution of steep, soil-mantled landscapes. *Geology* **33,** 349–352 (2005).
17. Benda, L. The influence of debris flows on channels and valley floors in the Oregon Coast Range, U.S.A. *Earth Surf. Process. Landf.* **15,** 457–466 (1990).
18. Jerolmack, D. J. & Paola, C. Shredding of environmental signals by sediment transport. *Geophys. Res. Lett.* **37,** L19401 (2010).
19. Mills, H. H. Apparent increasing rates of stream incision in the eastern United States during the late Cenozoic. *Geology* **28,** 955–957 (2000).
20. Gardner, T. W., Jorgensen, D. W., Shuman, C. & Lemieux, C. R. Geomorphic and tectonic process rates: effects of measured time interval. *Geology* **15,** 259–261 (1987).
21. Whittaker, A. C., Cowie, P. A., Attal, M., Tucker, G. E. & Roberts, G. P. Bedrock channel adjustment to tectonic forcing: implications for predicting river incision rates. *Geology* **35,** 103–106 (2007).
22. Lamb, M. P. & Fonstad, M. A. Rapid formation of a modern bedrock canyon by a single flood event. *Nature Geosci.* **3,** 477–481 (2010).
23. Schaller, M. *et al.* Fluvial bedrock incision in the active mountain belt of Taiwan from in situ-produced cosmogenic nuclides. *Earth Surf. Process. Landf.* **30,** 955–971 (2005).
24. Blöthe, J. H. & Korup, O. Millennial lag times in the Himalayan sediment routing system. *Earth Planet. Sci. Lett.* **382,** 38–46 (2013).
25. Kirchner, J. W. *et al.* Mountain erosion over 10 yr, 10 k.y., and 10 m.y. time scales. *Geology* **29,** 591–594 (2001).
26. Turowski, J. M., Hovius, N., Wilson, A. & Horng, M.-J. Hydraulic geometry, river sediment and the definition of bedrock channels. *Geomorphology* **99,** 26–38 (2008).
27. Ouimet, W. B., Whipple, K. X., Royden, L. H., Sun, Z. & Chen, Z. The influence of large landslides on river incision in a transient landscape: eastern margin of the Tibetan Plateau (Sichuan, China). *Geol. Soc. Am. Bull.* **119,** 1462–1476 (2007).
28. Gani, N. D. S., Gani, M. R. & Abdelsalam, M. G. Blue Nile incision on the Ethiopian Plateau: pulsed plateau growth, Pliocene uplift, and hominin evolution. *GSA Today* **17** (9), 4–11 (2007).

29. Sadler, P. & Jerolmack, D. J. Scaling laws for aggradation, denudation and progradation rates: the case for time-scale invariance at sediment sources and sinks. In *Strata and Time: Probing the Gaps in Our Understanding* (eds Smith, D. & Burgess, P.) (Geological Society of London Special Publication, in the press).

Acknowledgements This study was supported in part by the National Science Foundation (EAR-1049889). We thank J. Kirchner, B. Crosby and K. Ferrier for discussion and comments that improved the manuscript.

Author Contributions N.J.F. and S.F. compiled the datasets and performed the statistical analyses, R.S. provided guidance on the theory of stochastic processes, and N.J.F. wrote the paper with input from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.J.F. (nfinnega@ucsc.edu).

METHODS

To test for a dependence of river incision rate on measurement interval, we compiled 14 bedrock river incision data sets, each of which spans at least one order of magnitude in time (Supplementary Table 1), encompassing a total of 155 independent measurements of river incision. First we describe the incision data sources. Then, we describe the statistical analyses performed on the data to constrain the relationship between cumulative bedrock incision and measurement interval for the 14 data sets.

We provide the location and citation for each of the 14 incision data sets used in the analysis. In addition, we describe how both landform ages and uncertainties and cumulative bedrock incision values and uncertainties were determined for each data set. Of the 14 data sets, we identified tectonically active settings as those regions with documented tectonically driven rock uplift. In addition, each of the tectonically active data sets was taken from a paper that measured river incision to constrain active tectonic processes. Areas identified as tectonically inactive did not meet these criteria.

Middle Gorge of the Indus river, Pakistan. We used the weighted average ages and corresponding uncertainties and elevations reported in table 2 of ref. 3. Elevation uncertainty was assigned according to section 3.1 of ref. 3. We ignored point 5 in table 2 of ref. 3, following the interpretation of the authors.

Clearwater river, Olympic Mountains, Washington (state), USA. We used the ages and age ranges from figure 4 of ref. 30 to define the ages and age uncertainties for terraces. We digitized the terrace elevations from figure 4 of ref. 30 and used the reported elevation uncertainties in table 1 of ref. 30.

Umpqua and Siletz rivers, Oregon Coast Range, Oregon, USA. We used the ages and elevations along with their corresponding uncertainties as reported in table 1 of ref. 31. We selected only the Siletz and Umpqua River data sets as these span the longest period of time.

Bidente and Musone rivers, Northern Apennines, Italy. We used the ages and elevations along with their corresponding uncertainties as reported in tables 2 and 4 of ref. 32.

Yellow river, Linxia basin, eastern Tibetan plateau margin, China. We used the ages and bedrock elevations reported in table 1 of ref. 33. We assume a 10% uncertainty for age and a 5 m uncertainty in elevation, in keeping with the largest observed uncertainties in other data sets.

Sierra Nevada caves, California, USA. We used the ages and elevations along with their corresponding uncertainties as reported in table 2 of ref. 34.

Trachyte creek, Colorado plateau, Arizona, USA. We used strath ages and corresponding uncertainties reported in table 3 of ref. 35. Strath elevations were obtained from table 4 of ref. 35. We assigned an uncertainty in elevation of 5 m.

Grand Canyon, Colorado river, Colorado plateau, Arizona, USA. To avoid comparing reaches with different rates of tectonic forcing, we selected only data identified by the authors of ref. 36 as part of the "Western [Fault] Block" of the Grand Canyon because it is the longest record presented. Ages and uncertainties as well as strath elevations were obtained from table 2 of ref. 36. Strath elevation uncertainties were estimated to be half of the maximum pool depth reported in table 2 of ref. 36.

Yellow river, Jinshaan canyon, Oordos plateau, China. Strath ages and uncertainties were obtained from table 1 of ref. 37. Terrace elevations were selected from the Qilangwo locale in table 2 of ref. 37. Uncertainties in elevation were assumed to be 5 m.

Grand Canyon of the Nile, Blue Nile river, Ethiopia[28]. Ages and elevations were digitized from figure 5 of ref. 28. Uncertainties of 10% were assumed for age, and 5 m uncertainties were assigned to elevations.

Mammoth cave, Green river, Kentucky, USA. Ages and uncertainties were obtained from table 2 of ref. 38. Elevations were obtained from table 3 of ref. 38. Uncertainties in elevation were assigned according to the Methods of ref. 38.

Cumberland river, Cumberland plateau, Kentucky and Tennessee, USA. Ages and uncertainties as well as elevations were obtained from table 1 of ref. 39. Uncertainties in elevation were assigned to be 2 m, because of the similarity of this study to ref. 38.

We performed a Monte Carlo error analysis for each data set, in which we calculated 3,000 linear fits between the logarithm of cumulative bedrock incision and the logarithm of measurement interval. The measurement interval for a river incision measurement is equivalent to the age of the landform used to constrain incision. This is because cumulative river incision for all data points is computed between the modern channel and the palaeo-channel. We do not calculate incision between dated landforms because such estimates do not represent statistically independent measurements of incision. For each model iteration, we assigned incision and age errors by selecting randomly from a normal error distribution for each data point with a standard deviation corresponding to the reported or estimated uncertainty in landform age and elevation. Slope and intercept distributions were created for each data set from the results of the Monte Carlo simulation to define power-law exponents relating cumulative incision and measurement interval, as well as corresponding uncertainties. Because our explanatory variable (measurement interval) has significant uncertainty, we use a total-least-squares regression method to quantify slope.

30. Wegmann, K. W. & Pazzaglia, F. J. Holocene strath terraces, climate change, and active tectonics: the Clearwater River basin, Olympic Peninsula, Washington State. *Geol. Soc. Am. Bull.* **114,** 731–744 (2002).
31. Personius, S. F. Late Quaternary stream incision and uplift in the forearc of the Cascadia subduction zone, western Oregon. *J. Geophys. Res.* **100,** 20193–120210 (1995).
32. Wegmann, K. W. & Pazzaglia, F. J. Late Quaternary fluvial terraces of the Romagna and Marche Apennines, Italy: climatic, lithologic, and tectonic controls on terrace genesis in an active orogen. *Quat. Sci. Rev.* **28,** 137–165 (2009).
33. Li, J.-J. *et al.* Magnetostratigraphic dating of river terraces: rapid and intermittent incision by the Yellow River of the northeastern margin of the Tibetan Plateau during the Quaternary. *J. Geophys. Res.* **102,** 10121–10132 (1997).
34. Stock, G. M., Anderson, R. S. & Finkel, R. C. Rates of erosion and topographic evolution of the Sierra Nevada, California, inferred from cosmogenic ^{26}Al and ^{10}Be concentrations. *Earth Surf. Process. Landf.* **30,** 985–1006 (2005).
35. Cook, K. L., Whipple, K. X., Heimsath, A. M. & Hanks, T. C. Rapid incision of the Colorado River in Glen Canyon — insights from channel profiles, local incision rates, and modeling of lithologic controls. *Earth Surf. Process. Landf.* **34,** 994–1010 (2009).
36. Karlstrom, K. E. *et al.* ^{40}Ar/^{39}Ar and field studies of Quaternary basalts in Grand Canyon and model for carving Grand Canyon: quantifying the interaction of river incision and normal faulting across the western edge of the Colorado Plateau. *Geol. Soc. Am. Bull.* **119,** 1283–1312 (2007).
37. Cheng, S., Deng, Q., Zhou, S. & Yang, G. Strath terraces of Jinshaan Canyon, Yellow River, and Quaternary tectonic movements of the Ordos Plateau, North China. *Terra Nova* **14,** 215–224 (2002).
38. Granger, D. E., Fabel, D. & Palmer, A. N. Pliocene–Pleistocene incision of the Green River, Kentucky, determined from radioactive decay of cosmogenic ^{26}Al and ^{10}Be in Mammoth Cave sediments. *Geol. Soc. Am. Bull.* **113,** 825–836 (2001).
39. Anthony, D. M. & Granger, D. E. A late Tertiary origin for multilevel caves along the western escarpment of the Cumberland Plateau, Tennessee and Kentucky, established by cosmogenic ^{26}Al and ^{10}Be. *J. Caves Karst Stud.* **66,** 46–55 (2004).

Extended Data Table 1 | Power-law fits, tectonic setting and landform type for each data set

Location	Reference	Cumulative incision versus measurement interval exponent (β)	+/- 1σ	Length of incision record (kyr)	Tectonic setting	Landform
Middle gorge of the Indus river, Pakistan	3	0.67	0.19	65	Active	Unpaired strath terraces
Clearwater river, Olympic Mountains, Washington (state), USA	30	0.49	0.12	150	Active	Paired and unpaired strath terraces
Umpqua river, Oregon Coast Range, Oregon, USA	31	1.16	0.27	125	Active	Unpaired strath terraces
Siletz river, Oregon Coast Range, Oregon, USA	31	0.69	0.07	42	Active	Unpaired strath terraces
Bidente river, Northern Apennines, Italy	32	0.76	0.04	800	Active	Paired and unpaired strath terraces
Musone river, Northern Apennines, Italy	32	0.65	0.07	775	Active	Paired and unpaired strath terraces
Yellow river, Linxia Basin, eastern Tibetan plateau margin, China	33	0.8	0.16	1,660	Active	Strath terrace
Sierra Nevada caves, California, USA	34	1.84	0.28	2,700	Inactive	Cave
Trachyte creek, Colorado plateau, Arizona, USA	35	0.88	0.24	267	Inactive	Paired and unpaired strath terraces
Grand Canyon, Colorado river, Colorado plateau, Arizona, USA	36	0.57	0.17	5,500	Inactive	Miscellaneous
Yellow river, Jinshaan canyon, Oordos plateau, China	37	0.67	0.15	1,410	Inactive	Strath terrace
Gorge of the Nile, Blue Nile river, Ethiopia	28	0.44	0.01	3,002	Inactive	Dated volcanic rocks
Mammoth cave, Green river, Kentucky, USA	38	0.79	0.06	3,360	Inactive	Cave
Cumberland river, Cumberland plateau, Kentucky and Tennessee, USA	39	0.57	0.15	5,680	Inactive	Cave

LETTER

doi:10.1038/nature12797

Amazon River carbon dioxide outgassing fuelled by wetlands

Gwenaël Abril[1,2], Jean-Michel Martinez[2], L. Felipe Artigas[3], Patricia Moreira-Turcq[2], Marc F. Benedetti[4], Luciana Vidal[5], Tarik Meziane[6], Jung-Hyun Kim[7], Marcelo C. Bernardes[8], Nicolas Savoye[1], Jonathan Deborde[1], Edivaldo Lima Souza[9], Patrick Albéric[10], Marcelo F. Landim de Souza[11] & Fabio Roland[5]

River systems connect the terrestrial biosphere, the atmosphere and the ocean in the global carbon cycle[1]. A recent estimate suggests that up to 3 petagrams of carbon per year could be emitted as carbon dioxide (CO_2) from global inland waters, offsetting the carbon uptake by terrestrial ecosystems[2]. It is generally assumed that inland waters emit carbon that has been previously fixed upstream by land plant photosynthesis, then transferred to soils, and subsequently transported downstream in run-off. But at the scale of entire drainage basins, the lateral carbon fluxes carried by small rivers upstream do not account for all of the CO_2 emitted from inundated areas downstream[3,4]. Three-quarters of the world's flooded land consists of temporary wetlands[5], but the contribution of these productive ecosystems[6] to the inland water carbon budget has been largely overlooked. Here we show that wetlands pump large amounts of atmospheric CO_2 into river waters in the floodplains of the central Amazon. Flooded forests and floating vegetation export large amounts of carbon to river waters and the dissolved CO_2 can be transported dozens to hundreds of kilometres downstream before being emitted. We estimate that Amazonian wetlands export half of their gross primary production to river waters as dissolved CO_2 and organic carbon, compared with only a few per cent of gross primary production exported in upland (not flooded) ecosystems[1,7]. Moreover, we suggest that wetland carbon export is potentially large enough to account for at least the 0.21 petagrams of carbon emitted per year as CO_2 from the central Amazon River and its floodplains[8]. Global carbon budgets should explicitly address temporary or vegetated flooded areas, because these ecosystems combine high aerial primary production with large, fast carbon export, potentially supporting a substantial fraction of CO_2 evasion from inland waters.

In the global carbon cycle, rivers act not only as vectors from land to ocean but also as significant sources of CO_2 to the atmosphere. The amount of carbon that leaves the terrestrial biosphere through inland waters is much larger than the amount that ultimately reaches the ocean[1]. There is a growing consensus that outgassed carbon from inland waters originates from land, where it was fixed by terrestrial plants, then recycled within soils and finally exported to surface waters[1,2]. In water, the respiratory destruction of terrestrial organic carbon predominates over photosynthetic production[9,10]. However, other sources such as soil and groundwater CO_2, which are particularly significant in small streams, or wetland carbon, are necessary to balance CO_2 evasion at the global scale[2,4]. Over the past few years, estimates of the outgassing flux have substantially increased, in parallel with the development of remote sensing tools able to capture flooding[2]. Of the maximum 15% of global land area occupied by water, 9% is seasonally flooded[5]. Temporary and vegetated waters are grouped under the generic term of wetlands. Wetlands are among the most productive ecosystems on Earth[6] and behave as net carbon sinks owing to their efficient storage of carbon in waterlogged soils[11]. Yet they have been considered atmospheric CO_2 sources in inland water inventories[2,8]. Therefore, a better understanding of the carbon flows through the boundaries of wetlands with uplands, the atmosphere and rivers is crucial to determine an accurate continental carbon budget.

In this study, we examined the central Amazon River–floodplain system, which comprises the open waters and wetlands located in the 1.77-million-km^2 reference quadrant of the central Amazon basin that have been characterized in detail with Synthetic Aperture Radar (SAR) imagery[12] (Fig. 1b). Approximately 14% of this quadrant is occupied by wetlands; that is, temporary or vegetated waters in the floodplain, which are 96% inundated during high-water periods and 26% inundated during

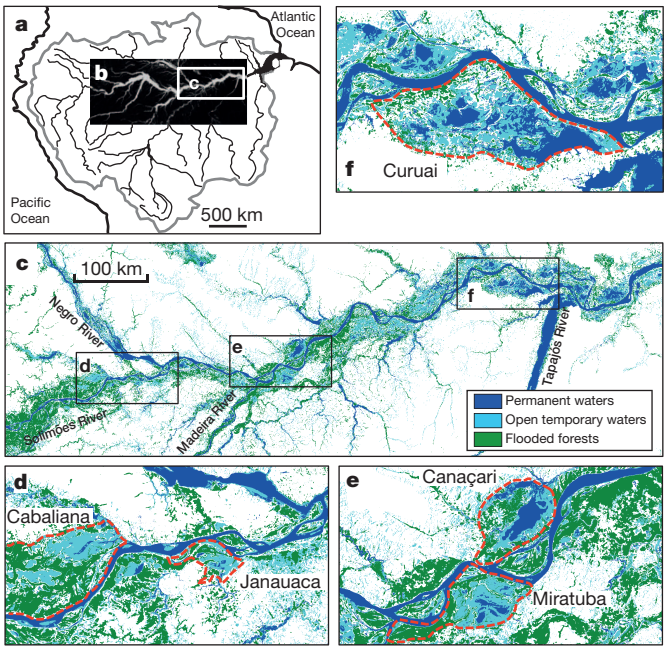


Figure 1 | Study area. a, The Amazon Basin. **b**, The Central Amazon reference quadrant (1.77 million km^2). **c**, The section of the Amazon River and the five floodplains investigated in this work. **d**, The Cabaliana and Janauaca floodplains on the Solimões river. **e**, The Canaçari and Miratuba floodplains on the Amazon River. **f**, The Curuai floodplain on the Amazon River.

[1]Laboratoire Environnements et Paléoenvironnements Océaniques et Continentaux (EPOC), CNRS, Université Bordeaux 1, Avenue des Facultés, 33405 Talence, France. [2]Laboratoire Géosciences et Environnement de Toulouse, Institut de Recherche pour le Développement, Université Paul Sabatier, 14 avenue Edouard Belin, 31400 Toulouse, France. [3]Laboratoire d'Océanologie et Géosciences, CNRS, Université du Littoral Côte d'Opale, 32 avenue Foch, 62930 Wimereux, France. [4]Equipe Géochimie des Eaux, Institut de Physique du Globe de Paris, Université Paris Diderot, Sorbonne Paris Cité, 35 rue Hélène Brion, 75205 Paris Cedex 13, France. [5]Laboratório de Ecologia Aquática, Departamento de Biologia, Universidade Federal de Juiz de Fora, Rua José Lourenço Kelmer, MG 36036-900 Juiz de Fora, Brazil. [6]Laboratoire Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), Muséum National d'Histoire Naturelle, CNRS, IRD, UPMC, 61 rue Buffon, 75005, Paris, France. [7]NIOZ (Royal Netherlands Institute for Sea Research), Department of Marine Organic Biogeochemistry, Texel, 1790 AB Den Burg, The Netherlands. [8]Programa de Geoquímica, Universidade Federal Fluminense, Outeiro São João Batista, RJ 24020015 Niterói, Brazil. [9]Instituto de Geociências, Universidade de Brasília, Campus Universitário Darcy Ribeiro, DF 70.910-900 Brasília, Brazil. [10]Institut des Sciences de la Terre d'Orléans, 1A rue de la Férollerie, 45071 Orléans Cedex 2, France. [11]Laboratório de Oceanografia Química, Universidade Estadual de Santa Cruz, Rodovia Ilhéus-Itabuna, 45662-900 Ilhéus, Bahia, Brazil.

low-water periods. Forest occupies about 70% of the entire flooded area during high-water periods[12]. These waters emit 210 ± 60 teragrams of carbon (Tg of C) per year as CO_2, of which approximately half is from the open waters and the other half is from the wetlands[8]. We combined high-resolution field measurements of the partial pressure of carbon dioxide (p_{co_2}) with remote sensing data on the extent of water and vegetation in the floodplain and mainstream. We focused on a section approximately 800 km long located in the farthest-downstream area in the reference quadrant (Fig. 1c). This section is of particular interest because the four characteristic biotopes of the central Amazon River (flooded forest, deep channels, shallow open floodplain lakes and floating macrophytes) are all well represented, although the proportions are different from those in the reference quadrant (Extended Data Table 1).

In addition, this section shows a well-defined biogeographic gradient from flooded forests that are dominant upstream to open lakes that are dominant downstream[12,13]. During eight periods from 2007–2011, we measured the water p_{co_2} continuously in the main channels of the Amazon River and its major tributaries and in five floodplain lakes (Fig. 1d–f). Our instrumental set-up was best suited to assess extreme spatial heterogeneity, including most of the remote and shallow water bodies (Extended Data Fig. 1). Except for some clear waters, such as the Tapajós River, the p_{co_2} in the main river channels was consistent with previous reports[8] and varied within one order of magnitude, from approximately 1,000 to 10,000 parts per million by volume (p.p.m.v.) (Extended Data Table 2). In floodplain lakes, the p_{co_2} varied by three orders of magnitude, from approximately 20 to 20,000 p.p.m.v. The

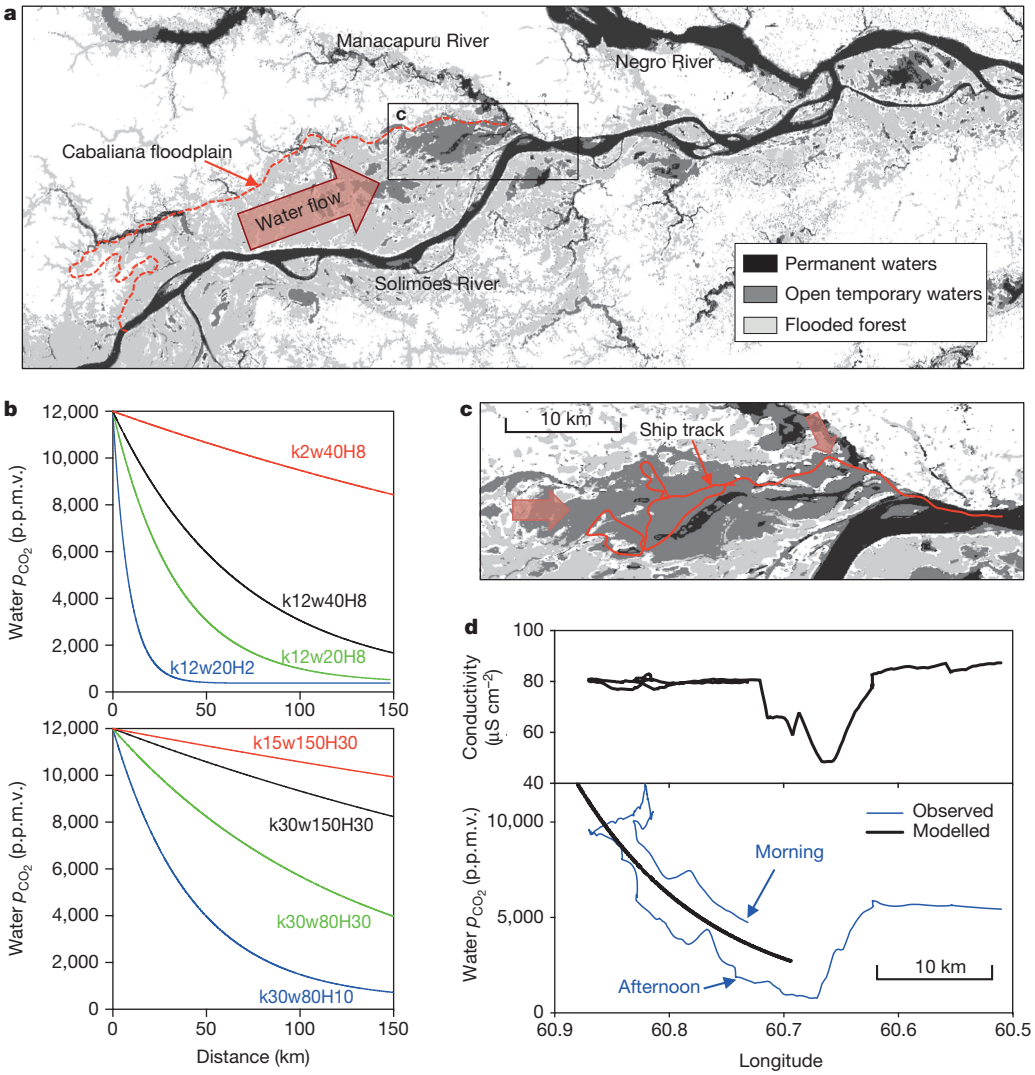


Figure 2 | CO_2 advection from flooded forest to open waters. **a**, The process of CO_2 advection could be tested using high-resolution p_{co_2} measurements performed in the Cabaliana Lake region, where waters have a single unidirectional eastward flow (from flooded forest (light grey) to open waters (dark grey and black)). **b**, Theoretical p_{co_2} distributions in floodplains (top graph) and river channels (bottom graph) calculated by the one-dimensional advection model (Supplementary Methods). The p_{co_2} curves start at a high p_{co_2} of 12,000 p.p.m.v. within wetlands and account for variable transport and degassing intensities. The models are labelled according to the parameters chosen: for example, model k2w40H8 indicates that the gas transfer velocity k is normalized to a Schmidt number of 600 and set to 2 cm h^{-1}, the water velocity w is set to 40 cm s^{-1}, and the water depth H is set to 8 m. Various settings for these three parameters corresponding to typical conditions in rivers and floodplains have been applied (Supplementary Information and Extended Data Fig. 5). **c**, The track of the ship in Cabaliana Lake and the Solimões river channel in May 2008. Brown arrows indicate the dominant water flows into the lake. **d**, Results of the measurements obtained while underway during the cruise. The conductivity distribution shows that waters from the Solimões river channel were dominant in the western part of Cabaliana Lake but were mixed with waters from the Manacapuru River in the eastern part of the lake. The water p_{co_2} showed a clear decrease from west to east as the distance from the flooded forest increased. The solid black line shows the theoretical p_{co_2} distribution modelled by considering a normalized gas transfer velocity of 12 cm h^{-1}, a water velocity of 10 cm s^{-1} and a water depth of 5 m, as was observed during the cruise. The differences in p_{co_2} at the centre of the lake in the morning and the afternoon were attributed to photosynthesis by phytoplankton, which were abundant in the lake.

extreme heterogeneity in the floodplains (Extended Data Figs 2 and 3) was related to the connections between the waters and vegetation at all spatial and temporal scales. At the centre of the open lakes, the water p_{CO_2} was highest during high-water periods in the floodplains surrounded by large flooded forests, whereas the lowest values occurred in almost totally isolated lakes during low-water periods and were associated with a high phytoplankton biomass (Extended Data Table 2). There was a net downstream decline in p_{CO_2} along the biogeographic gradient; the upstream forested floodplains and channels of the Solimões River always showed higher saturation values than the downstream open lakes and channels of the Amazon River. Finally, within individual lakes and during any season, p_{CO_2} increased consistently from open waters to the vicinity of floating macrophytes and flooded forests.

Wetland vegetation can support the outgassing of CO_2 in the surrounding water through two mechanisms, which differ in the form of carbon transported by the waters and in the site where respiration occurs (Extended Data Fig. 4). First, litterfall and root exudation release labile organic carbon to the water[14,15], where the organic carbon is further decomposed; in this organic carbon pathway, heterotrophic metabolism and outgassing occur concomitantly in the open waters. Second, submerged roots and microbial respiration in wetland soils release CO_2 to the water[16]; this CO_2 is then transported to open areas, where outgassing occurs. We investigated the potential magnitude of this latter CO_2 pathway with a one-dimensional model that computes the distance that wetland CO_2 is transported before it is emitted (Extended Data Fig. 5). We found that the water movement is fast enough relative to gas exchange to maintain high supersaturation over dozens to hundreds of kilometres without requiring heterotrophic metabolism. The ability of water masses to export wetland CO_2 depends on their velocity, depth, vertical mixing and the gas transfer coefficient k. In a river channel that is 30 m deep with a typical k value[8] of 15 cm h^{-1} and a water current of 150 cm s^{-1}, only 18% of the CO_2 has been evaded 150 km downstream of the point source (the rest is still dissolved in the river water and is transported further downstream). In contrast, in an open floodplain lake that is 2 m deep with a typical thermally enhanced k value[17] of 12 cm h^{-1} and a water current of 20 cm s^{-1}, 90% of the wetland CO_2 is outgassed 20 km downstream. Our model satisfactorily mimics the spatial p_{CO_2} gradient observed in May 2008 along a 20-km transect in Cabaliana Lake, where the water current flows permanently eastward after passing over a large flooded forest area (Fig. 2d).

The CO_2 outgassing fluxes in individual lakes increase with the percentage of the floodplain covered by vegetation (Fig. 3). Lakes that were almost isolated from wetlands had nearly neutral daily CO_2 fluxes, whereas lakes connected to large wetlands showed the highest outgassing rates. This relationship was valid in three different seasons for a given lake, which is consistent with the flood pulse concept[18], and also during the same season at different lakes because the vegetation cover decreases downstream. Floodplain vegetation that performs aerial photosynthesis thus drives CO_2 outgassing. The CO_2 pathway apparently dominates in floodplains (Fig. 2), but it also occurs in river channels, as indicated by the strong anomalies in the stoichiometry between dissolved CO_2 and oxygen concentrations. Indeed, respiration of inundated emergent vegetation can release dissolved CO_2 to waters through their roots with no concomitant drawdown of dissolved oxygen, which is supplied from the atmosphere through the leaves[16,19]. Net heterotrophy in lakes and main channels frequently accounts for less than 20% of the CO_2 outgassing flux[19,20], the majority of which is supported by wetlands CO_2 from upstream areas (Supplementary Information). Furthermore, the composition of particulate organic matter reveals a transfer of algal- and macrophyte-derived organic carbon from the floodplains to the river channel[21,22], and the isotopic composition of respired CO_2 in channels shows seasonally variable contributions from C3 plants, C4 plants and algal material[20,23], all of which are well represented in the floodplains. Flooded forest is also a potential major source of biodegradable macromolecules in the Amazon River[24]. Finally, both the organic carbon and CO_2 pathways are consistent with

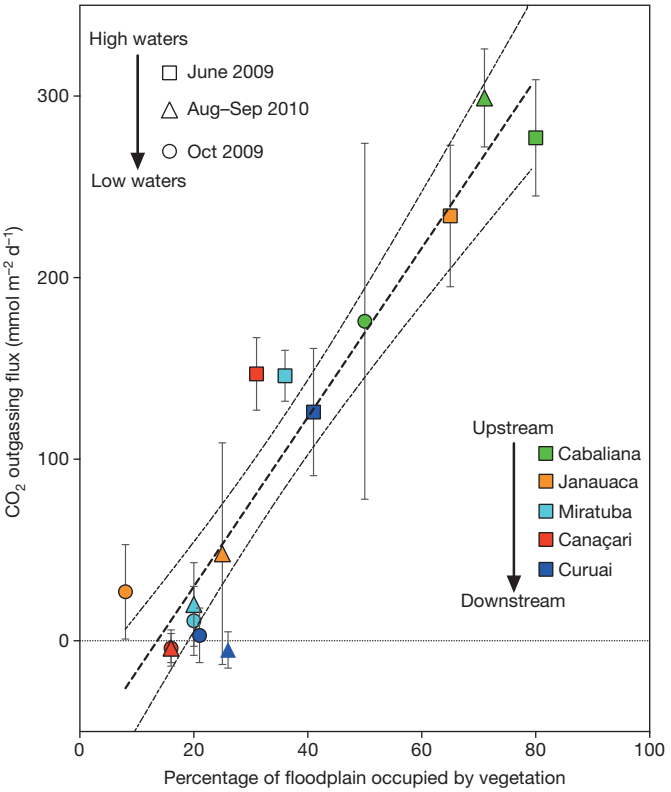


Figure 3 | Vegetation-mediated control of CO_2 outgassing in floodplain lakes. Symbols correspond to seasons, and colours correspond to sites. CO_2 degassing was computed from the measured p_{CO_2} and an average gas transfer velocity and was averaged over the surface area of the lake. Vertical bars indicate two standard deviations induced by the spatial and diurnal variation within each lake at a given season. The surface areas of the open lakes and the surrounding flooded and floating vegetation were obtained from SAR images taken during three seasons (Supplementary Methods). The thick dotted line represents the linear regression line ($r^2 = 0.90$, $P < 0.0001$); the thin dotted lines delineate the 95% confidence bands of the regression.

the very young ^{14}C age of the outgassed CO_2, compared with the older terrestrial organic carbon transported by the Amazon River[25].

Published metabolic rates[14–16,25,26] and the extent of vegetation[12] in the floodplains of the central Amazon reference quadrant allow some quantities of carbon supplied by wetlands to the waters to be computed. Flooded forests and floating macrophytes provide, through litterfall and submerged root respiration, a total of -305 ± 120 Tg of C per year of atmospheric carbon to the waters (Extended Data Table 3). This probably conservative estimate is not significantly different from the CO_2 outgassing flux of $+210 \pm 60$ Tg of C per year[8]. Central Amazonian waters thus receive at least as much carbon from semi-aquatic plants as they emit to the atmosphere, and the CO_2 net ecosystem exchange of the river system as a whole, including wetlands and open waters, should be nearly neutral. This is consistent with airborne observations over the Amazon, which do not reveal a large unidirectional CO_2 outgassing flux from wetland and river surfaces compared to uplands[27]. Water inside and downstream of a wetland functions similarly to the way soil functions in a forest, in that it rapidly returns a large fraction of the carbon fixed by photosynthesis to the atmosphere. Unlike in the headwaters[3], in the central Amazon, the contribution of strictly upland terrestrial carbon to CO_2 outgassing is potentially minor compared to the wetland carbon contribution. Unlike terrestrial landscapes that export less than 2% of their gross primary production to inland waters[1,7], Amazonian wetlands export half of their gross primary production to waters (Extended Data Table 3). Although more quantitative information is needed on the global distribution of wetland primary productivity and carbon export, the available information[28]

suggests that wetlands support a large percentage of inland water CO_2 evasion at the global scale.

METHODS SUMMARY

Continuous measurements (once per minute) were performed on eight cruises in the channels and floodplain lakes of the Amazon on a ship equipped for this purpose. River water was continuously pumped from about 50 cm below the surface and delivered to the different equipment and sensors. Water p_{co_2} was measured using a marble-type equilibrator[29] (see online Methods) connected to an infrared gas analyser, which was calibrated before and after each cruise. The response time was estimated to be 3 min. The water temperature, conductivity and turbidity were measured in an overflowing bucket using a multiprobe, which was calibrated every 10 days. Fluorescence was measured using a BBE Moldaenke Fluoroprobe. The position was recorded using a global positioning system (GPS). During low-water periods, shallow and remote lakes were surveyed using a small boat and a 12-V version of the measurement set-up. The total number of geo-referenced measurements was 45,786. The water and flooded vegetation surface areas were derived from SAR data acquired using the PALSAR space-borne sensor (see Supplementary Methods). Within the five floodplains sampled and for the three campaigns at high, intermediate and low water, the flooded areas were divided among open water, flooded forest and floating vegetation.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 February; accepted 25 October 2013.
Published online 15 December 2013.

1. Cole, J. J. *et al.* Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10,** 172–185 (2007).
2. Aufdenkampe, A. K. *et al.* Rivers key to coupling biogeochemical cycles between land, oceans and atmosphere. *Front. Ecol. Environ.* **9,** 53–60 (2011).
3. Davidson, E. A., Figueiredo, R. O., Markewitz, D. & Aufdenkampe, A. K. Dissolved CO_2 in small catchment streams of eastern Amazonia: a minor pathway of terrestrial carbon loss. *J. Geophys. Res.* **115,** G04005 (2010).
4. Butman, D. & Raymond, P. A. Significant efflux of carbon dioxide from streams and rivers in the United States. *Nature Geosci.* **4,** 839–842 (2011).
5. Downing, J. A. Global limnology: up-scaling aquatic services and processes to planet Earth. *Verh. Int. Verein. Limnol.* **30,** 1149–1166 (2009).
6. Whittaker, R. H. & Likens, G. E. in *Primary Productivity of the Biosphere* (eds Lieth, H. & Whittaker, R. H.) 305–328 (Springer, 1975).
7. Schulze, E. D. *et al.* The European carbon balance. Part 4: integration of carbon and other trace-gas fluxes. *Glob. Change Biol.* **16,** 1451–1469 (2010).
8. Richey, J. E., Melack, J. M., Aufdenkampe, A. K., Ballester, V. M. & Hess, L. L. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. *Nature* **416,** 617–620 (2002).
9. Battin, T. J. *et al.* Biophysical controls on organic carbon fluxes in fluvial networks. *Nature Geosci.* **1,** 95–100 (2008).
10. Duarte, C. M. & Prairie, Y. T. Prevalence of heterotrophy and atmospheric CO_2 emissions from aquatic ecosystems. *Ecosystems* **8,** 862–870 (2005).
11. Kayranli, B., Scholz, M., Mustafa, A. & Hedmark, A. Carbon storage and fluxes within freshwater wetlands: a critical review. *Wetlands* **30,** 111–124 (2010).
12. Hess, L. L., Melack, J. M., Novo, E. M., Barbosa, C. C. F. & Gastil, M. Dual-season mapping of wetland inundation and vegetation for the central Amazon basin. *Remote Sens. Environ.* **87,** 404–428 (2003).
13. Maurice-Bourgoin, L. *et al.* Temporal dynamics of water and sediment exchanges between the Curuaí floodplain and the Amazon River, Brazil. *J. Hydrol.* **335,** 140–156 (2007).
14. Schöngart, J., Wittmann, F. & Worbes, M. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 347–388 (Springer, 2010).
15. Engle, D. L., Melack, J. M., Doyle, R. D. & Fisher, T. R. High rates of net primary production and turnover of floating grasses on the Amazon floodplain: implications for aquatic respiration and regional CO_2 flux. *Glob. Change Biol.* **14,** 369–381 (2008).
16. Hamilton, S. K., Sippel, S. J. & Melack, J. M. Oxygen depletion and carbon dioxide and methane production in waters of the Pantanal wetland of Brazil. *Biogeochemistry* **30,** 115–141 (1995).
17. Polsenaere, P. *et al.* Thermal enhancement of gas transfer velocity of CO_2 in an Amazon floodplain lake revealed by eddy covariance. *Geophys. Res. Lett.* **40,** 1734–1740 (2013).
18. Junk, W. J., Bayley, P. B. & Sparks, R. E. The flood pulse concept in river–floodplain systems. in *Proc. Int. Large River Symp.* (ed. Dodge, D. P.) *Can. J. Fish. Aquat. Sci. Spec. Publ.* **106,** 110–127 (1989).
19. Devol, A. H. *et al.* Seasonal variation in chemical distributions in the Amazon (Solimões) River: a multiyear time series. *Glob. Biogeochem. Cycles* **9,** 307–328 (1995).
20. Ellis, E. E. *et al.* Factors controlling water-column respiration in rivers of the central and southwestern Amazon Basin. *Limnol. Oceanogr.* **57,** 527–540 (2012).
21. Mortillaro, J. M. *et al.* Particulate organic matter distribution along the Lower Amazon River: addressing aquatic ecology concepts using fatty acids. *PLoS ONE* **7,** e46141 (2012).
22. Moreira-Turcq, P. *et al.* Seasonal variability in concentration, composition, age and fluxes of particulate organic carbon exchanged between the floodplain and Amazon River. *Glob. Biogeochem. Cycles* **27,** 119–130 (2013).
23. Quay, P. D. *et al.* Carbon cycling in the Amazon River: implications from the ^{13}C compositions of particles and solutes. *Limnol. Oceanogr.* **37,** 857–871 (1992).
24. Ward, N. D. *et al.* Degradation of terrestrially derived macromolecules in the Amazon River. *Nature Geosci.* **6,** 530–533 (2013).
25. Mayorga, E. *et al.* Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. *Nature* **436,** 538–541 (2005).
26. Worbes, M. in *The Central Amazon Floodplain: Ecology of a Pulsing System* (ed. Junk, W. J.) 223–265 (Springer, 1997).
27. Lloyd, J. *et al.* An airborne regional carbon balance of Central Amazonia. *Biogeosciences* **4,** 759–768 (2007).
28. Aselmann, I. & Crutzen, P. J. Global distribution of natural freshwater wetlands and rice paddies, their net primary productivity, seasonality and possible methane emissions. *J. Atmos. Chem.* **8,** 307–358 (1989).
29. Abril, G., Richard, S. & Guérin, F. In situ measurements of dissolved gases (CO_2 and CH_4) in a wide range of concentrations in a tropical reservoir using an equilibrator. *Sci. Total Environ.* **354,** 246–251 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research is a contribution to the CARBAMA project, funded by the French National Agency for Research (grant number 08-BLAN-0221), the French INSU national programme EC2CO, and the National Council of Research and Development (CNPq), Brazil (Universal Program number 477655/2010-6). It was conducted under the auspices of the Environmental Research Observatory Hydrology and Geochemistry of the Amazon Basin (HYBAM), supported by the INSU and the IRD (Institute for Research and Development, France). F.R. was supported by CNPq and a Brazilian 'Excellent Researcher' fellowship. We thank all the participants of the CARBAMA cruises.

Author Contributions G.A., J.-M.M., P.M.-T., L.F.A., T.M. and M.F.B. conceived and designed the study. G.A. coordinated project and fieldwork. G.A., J.D., M.F.L.d.S. and N.S. performed the p_{co_2} measurements. J.-M.M. and E.L.S. analysed the remote sensing data. L.F.A. measured Chl *a* and fluorescence. L.V. and F.R. measured respiration. All authors contributed to the interpretation of the data. G.A. wrote the manuscript, J.-M.M., L.F.A. and F.R. contributed to manuscript writing and P.M.-T., L.V., T.M., J.-H.K., M.C.B., N.S. and M.F.B. commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.A. (g.abril@epoc.u-bordeaux1.fr).

METHODS

Study site and remote sensing. The central Amazon river–floodplain system is defined as the waters and wetlands located in the 1.77-million-km^2 reference quadrant of the central Amazon basin that has been characterized in detail using SAR imagery[12] (referred to as 'quadrant b'; Fig. 1b). We focused our fieldwork and remote sensing analysis on a secondary 0.288-million-km^2 quadrant inside the eastern limit of quadrant b, which was referred to as 'quadrant c' (Fig. 1c). Quadrant c is located in the downstream section of the Amazon River in Brazil, from the city of Manacapuru on the Solimões River upstream of its confluence with the Negro River to the Amazon River in front of the city of Santarem at the confluence of the Amazon with the Tapajós River (Fig. 1c). In this region, waters flowing from the Andes are rich in suspended matter and nutrients, have the highest conductivity and are termed 'white'[30]. Waters draining the primarily lowland podzol soils are rich in dissolved organic matter and intensely coloured by humic material, have low conductivity and are called 'black'. Finally, relatively transparent waters draining the Precambrian Shields are poor in organic and inorganic material, have low-to-intermediate conductivity and are called 'clear'. The main tributaries of the Amazon River sampled were the Solimões and Madeira (white), the Manacapuru and Negro (black), and the Urubu and Tapajós (clear). The Amazon River flood is characterized by relatively low interannual variability but strong water level differences between the low-water (October–November) and high-water stages (May–June). The annual water level within the main stream reaches approximately 8 m near Santarem station and up to 12 m at Manacapuru station on the Solimões River. The floodplains are connected to the river's main stem by a complex system of channels (called Paranás) that vary as a function of the flood level[31]. Spilling over the riverbanks usually occurs during the peak inundation phase[31,32]. These environments participate significantly in the transfer of water and matter from the Andes to the Atlantic Ocean; approximately 5% of the annual water discharge of the Amazon is estimated to travel through the floodplain[33].

Five floodplain systems were selected for fieldwork and remote sensing (Fig. 1 and Extended Data Table 1c). Farthest upstream, the Cabaliana and Janauacá floodplains receive large inputs of waters from the Solimões River and are surrounded by large areas of flooded forest (Fig. 1c, d). In the Cabaliana, the conductivity is similar to that of the Solimões (70–100 μS cm^{-1}), but the turbidity is much lower because of sediment retention in the flooded forest that acts as a filter. In the Janauacá, the conductivity is slightly lower (33–70 μS cm^{-1}) because of clear water contributed by local drainage. Downstream of the Amazon–Madeira confluence, the Canaçari and Miratuba mainly receive waters from the flooding of the Amazon and Madeira rivers, respectively (Fig. 1e). Miratuba Lake is well surrounded by flooded forest and is almost completely dry at low water (Fig. 1e). The Canaçari is a very homogeneous lake with approximately the same depth all over, and is disconnected from flooded forest most of the year. The Canaçari receives white waters from the Amazon and occasionally clear waters from the Urubu River that drains the northern local basin. Finally, in the farthest-downstream part of our study area, the Várzea of Curuai is composed of more than 30 interconnected lakes temporarily or permanently connected to the Amazon main stem by small channels[34] (Fig. 1f).

We used L-band SAR radar remote sensing images to monitor the extent of flooding at different stages in the five floodplains from 2007 to 2010. Several studies using a time series of L-band SAR satellite images reported excellent classification results in the quadrant under study[12,35,36]. The mapping accuracy was shown to increase with the number of images included in the time series (up to eight images[36]), which allowed the flood dynamics to be retrieved with an accuracy higher than 80%. The satellite images were acquired by the PALSAR space-borne sensor in wide-beam mode with 250-km-wide images and a resolution of 100 m. The PALSAR resolution allowed us to quantify the area of large lakes and flooded forest, but not small channels, in the floodplain. Because the area studied in this work extends over more than 800 km from east to west, five PALSAR wide-beam frames and eight images per frame were considered. The water-level range corresponding to the image acquisition agrees well with the maximum water level at the Obidos water gauge on the Amazon River, from 1.29 m to 8.56 m, and at the Manacapuru water gauge on the Solimões River, from 8.41 m to 20.43 m.

Flood maps were derived using an algorithm that allows the retrieval of the flood dynamics and the primary landscape units[36] with an accuracy better than 80% considering all landscape units and better than 90% for the forested areas. All images were co-registered and ground-corrected using the Shuttle Radar Topography Mission digital elevation model (http://srtm.csi.cgiar.org/) and georeferenced with MapReady software, available from the Alaska Satellite Facility. Speckle reduction was achieved by both temporal and spatial filtering using a specific algorithm for SAR image time series[37] and the Lee-Sigma filter[38]. Filtered multitemporal images were segmented into homogeneous regions using the multi-resolution segmentation algorithm of the eCognition software (http://www.ecognition.com/). The scale and shape parameter values that control the merging procedure were tested and applied uniformly to preserve the main floodplain landscape units. Then, a parallelepiped classifier was used to attribute a status regarding flooding to each segment: never flooded (NF), occasionally flooded (OF), and permanently flooded (PF). The classifier also provides the general type of land cover: open water, bare soil, savannah, forest. Combining all the images and scanning the time series from the lowest water level to the highest water level allowed the flooded surface areas to be retrieved at the minimum, maximum and intermediate flood levels for each landscape unit. The area flooded was assessed for each floodplain system to provide the variation in the areas of the flooded forests, the floating vegetation that emerges at the water surface, and the open waters of the floodplain lakes.

In quadrant c (Fig. 1), 31,435 km^2 are seasonally flooded, of which 37.7% is occupied by floodable forests, 10.1% by floating macrophytes, 17.6% is occupied by permanent river channels, 13.1% is permanent open lakes in the floodplain, and 21.1% is temporary open waters in the floodplain. Compared to the central Amazon reference quadrant (quadrant b in Fig. 1), open waters in the floodplain comprise more of our study area (Extended Data Table 1a, b). This sampling area thus provides a large array of floodplain conditions, from predominantly flooded forest floodplains, as in most of quadrant b, to open lakes that are more disconnected from wetland vegetation. There is a clear gradient in floodplain vegetation along this river section, beginning from the dominant flooded-forest coverage upstream on the Solimões River, which represents up to 75% of the total flooded area, and ending with the open-lake coverage dominant downstream on the Amazon River[13], in which only 10% of the floodplain is occupied by forest. The Curuai floodplain (2,333 km^2) and the Cabaliana floodplain (1,822 km^2) are the largest wetlands, whereas the Janauaca floodplain extends over only 216 km^2 at high water. The five floodplains studied show a downstream-decreasing trend of flooded-forest coverage (Extended Data Table 1c).

Field work. The CARBAMA database is the largest data set of water p_{co_2} and ancillary parameters (temperature, conductivity, dissolved oxygen, fluorescence and turbidity) recorded in a large tropical river–floodplain system. The total number of p_{co_2} measurements obtained was 47,786 (Extended Data Table 2), which, with a frequency of one measurement per minute, represents 32 days of continuous measurements while underway in the Amazon River, tributaries and five floodplain lakes. The eight cruises were distributed along the Amazon hydrograph to sample low water (November 2007, October 2008 and October 2009), rising water (January 2007 and January 2011), high water (May 2008 and June 2009), and falling water (August–September 2010).

The cruises consisted of continuous measurements while underway (Extended Data Fig. 1), as well as discrete water sampling for chlorophyll *a* (Chl *a*) analysis. The continuous measurement system consisted of a circuit of river water installed in a typical Amazonian ship, with a bypass at the engine cooling water inlet located at a depth of about 40 cm. This water flowed gravimetrically into a five-litre bucket located in the vessel's hold that contained a submerged 220-V pump that raised water 5 m at a rate of 8 litres per minute. The water was allowed to overflow the bucket, and the inflow was optimized to exclude gas bubbles and reduce water residence time in the bucket, even when the ship was sailing at maximum speed, around 20 km h^{-1}. The pump pushed water from the ship's hold to a one-litre bucket, which was also continuously overflowing, located on the side of the boat. The water volume in this second bucket at deck level was minimized because it contained an Yellow Spring Instruments (YSI) multisensory probe that recorded ancillary parameters and a 12-V submerged pump that delivered water at a constant flow to the equilibrator (about one litre per minute) and to a closed circuit passing through a BBE-Moldaenke fluorimeter (2–3 litres per minute). There was a height difference of approximately 2 m between the first bucket in the hold and the container with the sensors; the equilibrator was placed approximately 1 m above the deck floor. By adding salt to the first bucket and measuring the conductivity in the second bucket, we determined the maximum water residence time to be 2.5 min. Comparison with direct measurements in the river revealed that the water temperature increased by no more than 0.5 °C in the circuit. There was no notable difference in the oxygen concentrations measured with the same optode inside and outside the water circuit. A Garmin global positioning system recorded the position every minute. The instrumental set-up was also converted to a 12-V version that could be used on small boats able to enter large, remote shallow lakes inaccessible to vessels at low water levels (Extended Data Fig. 1). In this case, the 12-V pump and the YSI probes were submerged directly from the side of the boat.

In a marble-type equilibrator, water flows from the top to the bottom (0.7–1 litre per minute), while a constant volume of air (about 0.5 litres per minute) flows from the bottom to the top; surface gas exchange is optimized around the marbles, and the volumes of water and air are minimized by the presence of the marbles. p_{co_2} in the air circuit tends rapidly to equilibrate to the water p_{co_2} value[29,39]. The pump delivers air from the top of the equilibrator to a LI820 LICOR gas analyser, and then back to the bottom of the equilibrator. The air from the equilibrator passes

first through a water trap, then through a desiccation tube (10 cm long, 2 cm in diameter) containing a desiccant (Drierite), then through the air pump, the gas analyser, a gas flow meter, and finally, back to the bottom of the equilibrator[29,39]. All tubes are made of Tygon or polytetrafluoroethylene (both gas-tight plastics) and the connections are also gas-tight. If the equilibrator is bypassed, the air circuit loses less than 1 p.p.m.v. of CO_2 per minute at about 3,000 p.p.m.v. The drying tube was intentionally small to limit the effect of the interaction of Drierite with CO_2 (adsorption/desorption), which increases the response time of the system[40]. We changed the Drierite in the tube every day and flushed the system with atmosphere for half an hour before measurements. The Drierite was regularly regenerated on board using a hairdryer. In the laboratory, before and after each cruise, the LICOR gas analyser was calibrated in the 0–20,000 p.p.m.v. range by nitrogen passing through soda lime to calibrate the zero and by a gas standard certified at 5,000 ± 30 p.p.m.v. (Air Liquide) to calibrate the span. The linearity was checked using another standard at 500 ± 8 p.p.m.v. During cruises, as advised by the manufacturer, we regularly checked the zero of the instrument by passing air over soda lime, and we reset the zero when the deviation was more than ±10 p.p.m.v. After the cruise, the maximum deviation with the 5,000 p.p.m.v. standard was ± 90 p.p.m.v., that is, less than 2%.

The YSI multiprobe measured the water temperature, conductivity, dissolved oxygen, pH and turbidity. We applied the calibration procedures recommended by the manufacturer. Calibration was performed at the beginning of each cruise and at the middle of cruises that lasted more than 10 days. The BBE-Moldaenke Fluoroprobe is a submersible spectral fluorometer that allows continuous recording of total *in vivo* fluorescence and discriminates among the main spectral groups of phytoplankton (that is, diatoms and dinoflagellates, blue-green algae, cryptophytes and green algae) on the basis of the relative fluorescence intensity of Chl *a* at 680 nm (because of the Photosystem II core pigments) after sequential excitation of the antenna and accessory pigments by five light-emitting diodes at 450 nm, 525 nm, 570 nm, 590 nm and 610 nm (refs 41, 42). The Chl *a* concentration was measured in two to three discrete water samples collected in each floodplain and river. The samples were filtered through Whatman 47-mm GF/F glass-fibre filters, stored frozen (in liquid nitrogen) until the return to the laboratory, and then extracted in 90% acetone overnight. Fluorescence measurements were performed using a Turner Designs 10-AU fluorometer before and after acidification with hydrochloric acid[43].

Respiration in river and lake surface waters was measured with 24-h incubations. Fifteen calibrated 60-ml Biochemical Oxygen Demand (BOD) bottles were filled with a well-homogenized and re-oxygenated (about 80% saturated) water sample. Five bottles were used for time zero, and ten were incubated in the dark at the temperature of the river water. Five bottles were fixed at 12 h, and the last five bottles were fixed at 24 h. In floodplain lakes, where phytoplanktonic production may occur, oxygen production in the light was measured by applying the same procedure, except that the samples were incubated on-deck at incident light, in a bucket thermostated with lake water. The oxygen concentrations were measured onboard the vessel with Winkler titration. Respiration and production were calculated from the linear regression of the oxygen concentration as a function of time; the correlation coefficient was greater than 0.90. The respiration rates were multiplied by the water depth to estimate the depth-integrated respiration, with an assumption of vertical homogeneity[20]. In the case of turbid river channels, where water depth is high and light penetration is very low, we assume that phytoplanktonic production does not occur and that the net heterotrophic metabolism equals the depth-integrated respiration[20]. With floodplain lakes, we made this assumption only for the high-water period, when phytoplankton biomass was low.

Flux computation. The air-water flux of CO_2 was calculated from the water temperature and p_{co_2} according to:

$$F(CO_2) = k\alpha\Delta p_{co_2}$$

where k is the gas transfer velocity of CO_2 (in units of cm h^{-1} or m s^{-1}), α is the solubility coefficient of CO_2 (in mol kg^{-1} atm^{-1}), and Δp_{co_2} is the air–water gradient of p_{co_2} (in atm). The atmospheric p_{co_2} measured every day at the beginning and end of the equilibrator records was 402 ± 24 p.p.m.v. and did not significantly change with the seasons, between morning and evening, or between lakes and channel sites; this average atmospheric value was used for all flux calculations, and Δp_{co_2} was largely driven by fluctuation in the water p_{co_2}. We used the polynomial formulation that gives the solubility coefficient α as a function of water temperature[44]. The gas transfer velocity k was calculated from the gas transfer velocity normalized to a Schmidt number of 600 that corresponds to CO_2 at 20 °C (ref. 45):

$$k = k_{600}\,(600/\mathrm{Sc})^n$$

where k_{600} is the normalized gas transfer velocity, Sc is the Schmidt number of a given gas at a given temperature[46], and n equals 2/3 in floodplain lakes and 0.5 in the more turbulent river channels[47].

Despite numerous studies on the subject, important uncertainty still remains regarding the parameterization of k_{600} in lakes, particularly in Amazon floodplain lakes. The widely used parameterizations based on wind speed[47,48] do not account for other drivers (or inhibitors) of turbulence at the water–air interface, such as the friction of the current on the bottom[49,50], evaporative heat gain and loss and the associated buoyancy fluxes[51,52], or the release of organic molecules that might behave as surfactants[53] by wetland vegetation[54]. Recently, several authors[52,55] have hypothesized that the k_{600} value of 2.7 cm h^{-1} used in previous estimates of CO_2 outgassing from the Amazon floodplains[8] is underestimated by a factor of four to five. Experimental data of k_{600} based on eddy covariance[17] confirm this underestimation exclusively in open lakes under high light incidence and heat flux; on the contrary, the k_{600} values may be much lower within a flooded forest, where no k_{600} data are available, where waters are protected from heat and wind by the canopy, where organic surfactants may be particularly abundant[54] and which represent 70% of the floodable area in the Amazon[12]. Because of the paucity of reliable information on k_{600} in floodplains, we provide average CO_2 fluxes from floodplains based on the value of 2.7 cm h^{-1}, and this value has the advantage of allowing quantitative comparison with the CO_2 fluxes reported previously[8]. When a more precise k_{600} value seems necessary, for instance, in the quantitative comparison between respiration rates and outgassing fluxes, or for the calculation of the distance of wetland CO_2 export, we apply the Cole and Caraco relationship[48] to the wind speeds at 10 m measured on the roof of the vessel at stations during the cruise. For the river channels, we use the value of 8.2 cm h^{-1} used by Richey *et al.*[8] and based on direct measurements of O_2 and ^{222}Rn accumulation in free-floating chambers[56,57]. In all cases, the chosen k_{600} values are specified in the text. It is worthwhile to note that none of the main conclusions of our paper is affected by the choice of the k_{600} value.

The revised carbon budget was computed at the scale of quadrant b, multiplying published metabolic rates[14–16,26,58–70] by their respective surface area (Extended Data Table 3).

30. Sioli, H. Hydrochemistry and geology in the Brazilian Amazon region. *Amazoniana* **3,** 267–277 (1968).
31. Mertes, L. A. K., Dunne, T. & Martinelli, L. A. Channel–floodplain geomorphology along the Solimões-Amazon River, Brazil. *Geol. Soc. Am. Bull.* **108,** 1089–1107 (1996).
32. Trigg, M. A., Bates, P. D., Wilson, M. D., Schumann, G. & Baugh, C. Floodplain channel morphology and networks of the middle Amazon River. *Wat. Resour. Res.* **48,** W10504 (2012).
33. Alsdorf, D., Han, S.-C., Bates, P. & Melack, J. Seasonal water storage on the Amazon floodplain measured from satellites. *Remote Sens. Environ.* **114,** 2448–2456 (2010).
34. Bonnet, M. P. *et al.* Floodplain hydrology in an Amazon floodplain lake (Lago Grande de Curuai). *J. Hydrol.* **349,** 18–30 (2008).
35. Rosenqvist, A., Forsberg, B. R., Pimentel, T., Rauste, Y. A. & Richey, J. E. The use of spaceborne radar to model inundation patterns and trace gas emissions in the central Amazon floodplain. *Int. J. Remote Sens.* **23,** 1303–1328 (2002).
36. Martinez, J. M. & Le Toan, T. Mapping of flood dynamics and spatial distribution of vegetation in the Amazon floodplain using multitemporal SAR data. *Remote Sens. Environ.* **108,** 209–223 (2007).
37. Quegan, S., Le Toan, T., Yu, J. J., Ribbes, F. & Floury, N. Multitemporal ERS SAR analysis applied to forest monitoring. *IEEE Trans. Geosci. Rem. Sens.* **38,** 741–753 (2000).
38. Lee, J. S. A simple speckle smoothing algorithm for synthetic aperture radar images. *IEEE Trans. Syst. Man Cybern.* **13,** 85–89 (1983).
39. Frankignoulle, M., Borges, A. & Biondo, R. A new design of equilibrator to monitor carbon dioxide in highly dynamic and turbid environments. *Water Res.* **35,** 1344–1347 (2001).
40. Santos, I. R., Maher, D. T. & Eyre, B. D. Coupling automated radon and carbon dioxide measurements in coastal waters. *Environ. Sci. Technol.* **46,** 7685–7691 (2012).
41. Beutler, M. *et al.* A fluorometric method for the differentiation of algal populations *in vivo* and *in situ*. *Photosynth. Res.* **72,** 39–53 (2002).
42. MacIntyre, H. L., Lawrenz, E. & Richardson, T. L. in *Chlorophyll a Fluorescence in Aquatic Sciences: Methods and Applications* (eds Suggett, D. J. *et al.*) 129–169 (Developments in Applied Phycology 4, Springer, 2010).
43. Lorenzen, C. J. Determination of chlorophyll and pheopigments: spectrophotometric equations. *Limnol. Oceanogr.* **12,** 343–346 (1967).
44. Weiss, R. F. Carbon dioxide in water and seawater: the solubility of a non-ideal gas. *Mar. Chem.* **2,** 203–215 (1974).
45. Jähne, B. *et al.* On parameters influencing air-water exchange. *J. Geophys. Res.* **92,** 1937–1949 (1987).
46. Wanninkhof, R. Relationship between gas exchange and wind speed over the ocean. *J. Geophys. Res.* **97,** 7373–7382 (1992).
47. Guérin, F. *et al.* Gas transfer velocities of CO_2 and CH_4 in a tropical reservoir and its river downstream. *J. Mar. Syst.* **66,** 161–172 (2007).

48. Cole, J. J. & Caraco, N. F. Atmospheric exchange of carbon dioxide in a low-wind oligotrophic lake measured by the addition of SF_6. *Limnol. Oceanogr.* **43,** 647–656 (1998).
49. Zappa, C. J. *et al.* Environmental turbulent mixing controls on air-water gas exchange in marine and aquatic systems. *Geophys. Res. Lett.* **34,** http://dx.doi.org/10.1029/2006GL028790 (2007).
50. Abril, G., Commarieu, M. V., Sottolichio, A., Bretel, P. & Guérin, F. Turbidity limits gas exchange in a large macrotidal estuary. *Estuar. Coast. Shelf Sci.* **83,** 342–348 (2009).
51. MacIntyre, S. *et al.* Buoyancy flux, turbulence, and the gas transfer coefficient in a stratified lake. *Geophys. Res. Lett.* **37,** L24604 (2010).
52. Rudorff, C. M., Melack, J. M., MacIntyre, S., Barbosa, C. C. F. & Novo, E. M. L. M. Seasonal and spatial variability of CO_2 emission from a large floodplain lake in the lower Amazon. *J. Geophys. Res.* **116,** G04007 (2011).
53. Salter, M. E. *et al.* Impact of an artificial surfactant release on air-sea gas fluxes during deep ocean gas exchange experiment II. *J. Geophys. Res.* **116,** C11016 (2011).
54. Parolin, P. *et al.* Central Amazon floodplain forests: tree survival in a pulsing system. *Bot. Rev.* **70,** 357–380 (2004).
55. Richey, J. E., Krusche, A. V., Johnson, M. S., da Cunha, H. B. & Ballester, M. V. in *Amazonia and Global Change* (eds Keller, M. *et al.*) 489–504 (Geophys. Monogr. Ser. 186, AGU, 2009).
56. Devol, A. H., Quay, P. D., Richey, J. E. & Martinelli, L. A. The role of gas exchange in the inorganic carbon, oxygen and ^{222}Rn budgets of the Amazon River. *Limnol. Oceanogr.* **32,** 235–248 (1987).
57. Alin, S. R. *et al.* Physical controls on carbon dioxide transfer velocity and flux in low-gradient river systems and implications for regional carbon budgets. *J. Geophys. Res.* **116,** G01009 (2011).
58. Junk, W. J. & Piedade, M. T. F. Biomass and primary production of herbaceous plant communities in the Amazon floodplain. *Hydrobiology* **263,** 155–162 (1993).
59. Malhi, Y. & Grace, J. Tropical forests and atmospheric carbon dioxide. *Trees* **15,** 332–337 (2000).
60. Horna, V., Zimmermann, R., Müller, E. & Parolin, P. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 223–241 (Springer, 2010).
61. Worbes, M. in *The Central Amazon Floodplain: Ecology of a Pulsing System* (ed Junk, W. J.) 223–265 (Springer, 1997).
62. Saatchi, S. S., Houghton, R. A., Dos Santos Avala, R. C., Soares, J. V. & Yu, Y. Distribution of aboveground live biomass in the Amazon basin. *Glob. Change Biol.* **13,** 816–837 (2007).
63. Meyer, U., Junk, W. J. & Linck, C. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 163–178 (Springer, 2010).
64. Parolin, P., Wittmann, F. & Schöngart, J. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 105–126 (Springer, 2010).
65. Piedade, M. T. F., Ferreira, C. S., de Oliveira Wittmann, A., Buckeridge, M. & Parolin, P. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 127–139 (Springer, 2010).
66. Junk, W. J., Piedade, M. T. F., Parolin, P., Wittmann, F. & Schöngart, J. in *Amazonian Floodplain Forests: Ecophysiology, Biodiversity and Sustainable Management* (eds Junk, W. J. *et al.*) 511–540 (Springer, 2010).
67. Morison, J. I. L. *et al.* Very high productivity of the C4 aquatic grass *Echinocloa polystachya* in the Amazon floodplain confirmed by net ecosystem CO_2 flux measurements. *Oecologia* **125,** 400–411 (2000).
68. Costa, M. Estimate of net primary productivity of aquatic vegetation of the Amazon floodplain using Radarsat and JERS-1. *Int. J. Remote Sens.* **26,** 4527–4536 (2005).
69. Melack, J. M. *et al.* in *Amazonia and Global Change* (eds Keller, M. *et al.*) 525–542 (Geophys. Monogr. Ser. 186, AGU, 2009).
70. Moreira-Turcq, P. *et al.* Carbon sedimentation at Lago Grande de Curuai, a floodplain lake in the low Amazon region: insights into sedimentation rates. *Palaeogeogr. Palaeoclim. Palaeoecol.* **214,** 27–40 (2004).

LETTER

doi:10.1038/nature12939

Upwash exploitation and downwash avoidance by flap phasing in ibis formation flight

Steven J. Portugal[1], Tatjana Y. Hubel[1], Johannes Fritz[2], Stefanie Heese[2], Daniela Trobe[2], Bernhard Voelkl[2,3]†, Stephen Hailes[1,4], Alan M. Wilson[1] & James R. Usherwood[1]

Many species travel in highly organized groups[1–3]. The most quoted function of these configurations is to reduce energy expenditure and enhance locomotor performance of individuals in the assemblage[4–11]. The distinctive V formation of bird flocks has long intrigued researchers and continues to attract both scientific and popular attention[4,7,9–14]. The well-held belief is that such aggregations give an energetic benefit for those birds that are flying behind and to one side of another bird through using the regions of upwash generated by the wings of the preceding bird[4,7,9–11], although a definitive account of the aerodynamic implications of these formations has remained elusive. Here we show that individuals of northern bald ibises (*Geronticus eremita*) flying in a V flock position themselves in aerodynamically optimum positions, in that they agree with theoretical aerodynamic predictions. Furthermore, we demonstrate that birds show wingtip path coherence when flying in V positions, flapping spatially in phase and thus enabling upwash capture to be maximized throughout the entire flap cycle. In contrast, when birds fly immediately behind another bird—in a streamwise position—there is no wingtip path coherence; the wing-beats are in spatial anti-phase. This could potentially reduce the adverse effects of downwash for the following bird. These aerodynamic accomplishments were previously not thought possible for birds because of the complex flight dynamics and sensory feedback that would be required to perform such a feat[12,14]. We conclude that the intricate mechanisms involved in V formation flight indicate awareness of the spatial wake structures of nearby flockmates, and remarkable ability either to sense or predict it. We suggest that birds in V formation have phasing strategies to cope with the dynamic wakes produced by flapping wings.

Theories of fixed-wing aerodynamics have predicted the exact spanwise positioning that birds should adopt in a V formation flock to

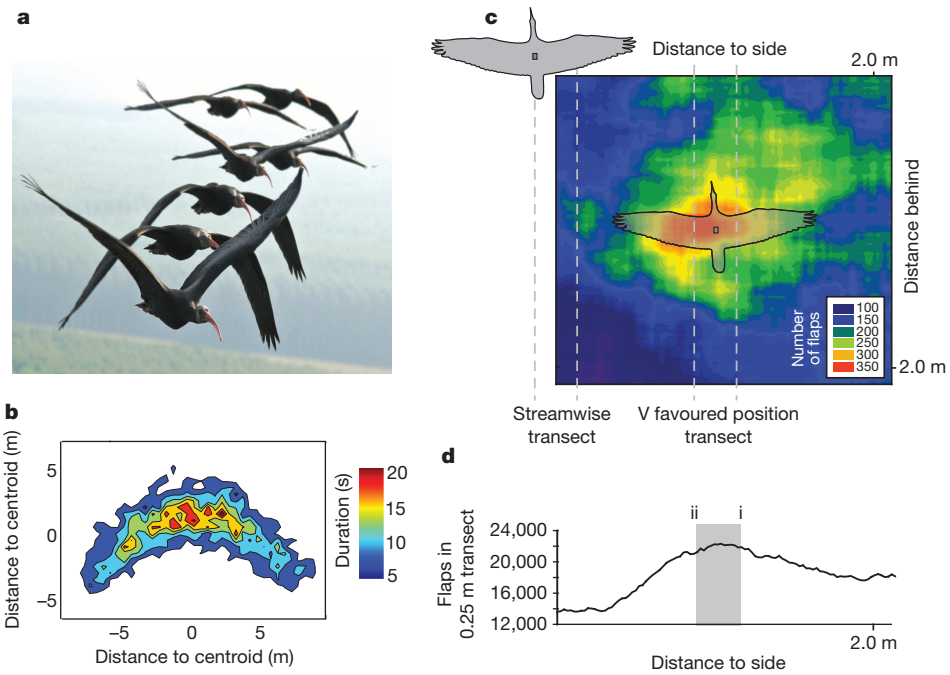


Figure 1 | V formation flight in migrating ibises. **a**, Northern bald ibises (*G. eremita*) flying in V formation during a human-led migratory flight (photograph M. Unsöld). **b**, Three-dimensional location histogram of the 7 min flight section, showing position of individual ibises ($n = 14$) in the V formation, with respect to flock centroid, measured by a 5 Hz GPS data logger. The colour scale refers to the duration (in seconds) a bird was present in each 0.25 m × 0.25 m grid. A plot detailing the formation shape for the duration of the entire flight can be found in Supplementary Fig. 7. **c**, Histogram of number of flaps (colour coded) recorded in each 0.25 m × 0.25 m region between all birds and all other birds. Most flaps occurred at an angle of approximately 45° to the bird ahead (or behind). Transects denoted by dashed lines, directly behind or along the most populated V favoured position (just inboard of wingtip to wingtip), are the same as those detailed in Fig. 3. **d**, Histogram detailing the total number of flaps recorded between each bird–bird pair, with respect to position of the following bird. The shaded area (ii–i) denotes the limits of optimal relative positioning, based on fixed-wing aerodynamics.

[1]Structure & Motion Laboratory, the Royal Veterinary College, University of London, Hatfield, Hertfordshire AL9 7TA, UK. [2]Waldrappteam, Schulgasse 28, 6162 Mutters, Austria. [3]Institute for Theoretical Biology, Humboldt University at Berlin, Invalidenstrasse 43, 10115 Berlin, Germany. [4]Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK. †Present address: Edward Grey Institute, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.

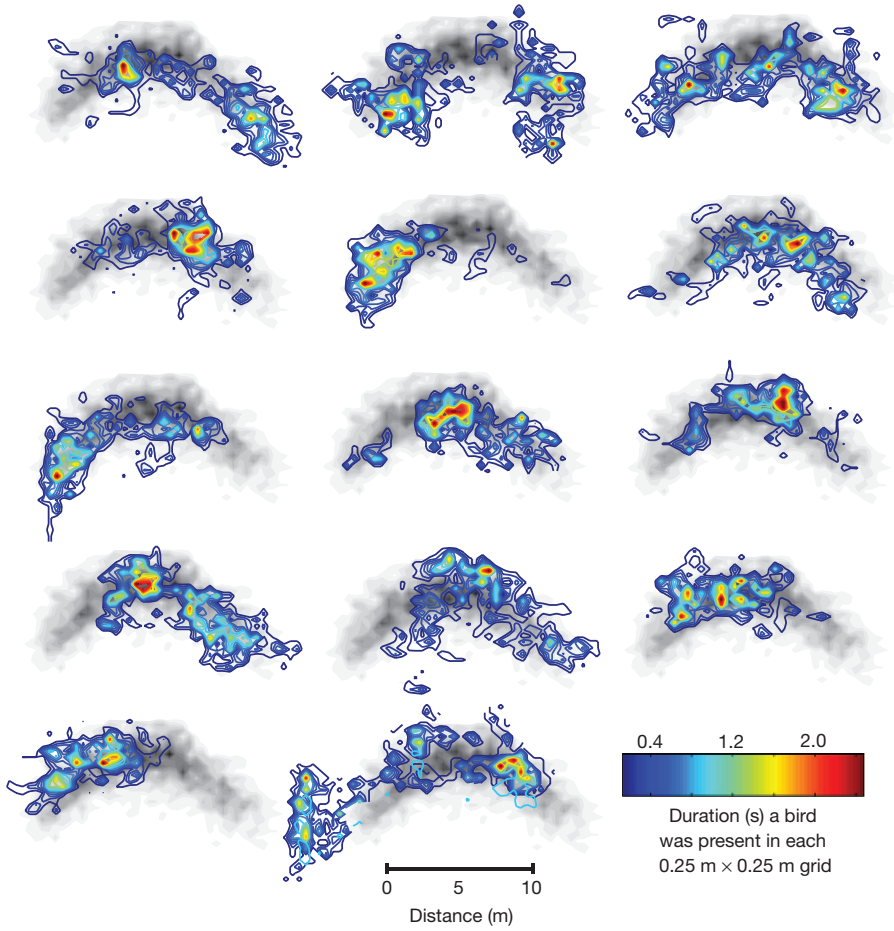


Figure 2 | Histograms demonstrating the positional infidelity for each northern bald ibis in the V formation during the migratory flight. The grey shaded V shape behind each individual histogram ($n = 14$) denotes the structure for all individuals in the flock (see Fig. 1b). The colour code refers to the duration (in seconds) a bird was present in each 0.25 m × 0.25 m grid. Although individual birds showed some bias towards the front, back, left or right regions of the V formation, these positions were not maintained rigidly.

maximize upwash capture[4,9–14]. The primary empirical evidence confirming that this mechanism is used is a reduction in heart rate and wing-beat frequency in pelicans flying in a V formation[7]. There is a general lack of experimental data from free-flying birds, mainly because of the complications of measuring the intricate and three-dimensional complexity of formation flight, and the lack of appropriate devices to monitor and record such information. Therefore, the precise aerodynamic interactions that birds use to exploit upwash capture have not been identified. To investigate the purported aerodynamic interactions of V formation flight, we studied a free-flying flock of northern bald ibises (*Geronticus eremita*) (Fig. 1a), a critically endangered migratory species. We used new technology[15,16] to measure the position, speed and heading of all birds in a V formation. We recorded position and every wing flap of 14 birds during 43 min of migratory flight using back-mounted integrated global positioning system (GPS) (5 Hz) and inertial measurement units (300 Hz) (see Methods)[15,16]. The precision of these measurements allows the relative positioning of individuals in a V to be tracked, and the potential aerodynamic interactions to be investigated at a level and complexity not previously feasible.

During a 7 min section of the flight, where most of the flock flew in approximate V formation in steady, level and planar direct flight (see Methods), we found wing flaps occurred at an angle of, on average, 45° to the bird ahead (or behind), and approximately 1.2 m behind (Fig. 1b–d). The most populated 1 m × 1 m region was 0.49–1.49 m behind and to the side of the bird ahead. The centre of the most populated (0.25 m) spanwise region was at 0.904 m, resulting in a wingtip overlap[9–13] of 0.115 m (Fig. 1c, d; wingspan $b = 1.2$ m). This falls within the bounds of predictions of fixed-wing theory[9–13] for maximizing the benefits from upwash, which range from zero wingtip overlap (assuming no wake contraction[4]) to, maximally, 0.13 m (assuming elliptical loading over the pair of wings, and full wake contraction from wingspan b to $\pi b/4$)[9].

During this 7 min section of V formation flight, individual birds show a certain degree of positional infidelity in the V flock (Fig. 2; see also Supplementary Fig. 1 and Supplementary Video 1). Although individuals contribute to the statistical V formation, their positioning is inconsistent. Certain individuals showed general preferences for a particular area in the V formation, but the variability in positioning in the flock resulted in no clear leader (see Supplementary Information for further discussion).

Although we observe that, when flying in a V, ibises position themselves in locations predicted mathematically from fixed-wing aerodynamics[4,9–11], the wake of flapping birds (in this study, ibises spent 97% of their time flapping; see Methods) is likely to be complex[9–14]. Wingtip path coherence, where a flying object flaps its wings in spatial phase with that of the individual it is following, has been proposed as a method that would maximize upwash capture in V formation flight of birds and flying robotic devices[12]. Whether birds are able to take advantage of this extra level of complexity present in flapping flight (compared with that of fixed-wing flight) has remained unanswered so far.

In the ibis flock, individual flaps for each bird were described from the dorsal acceleration signal from the inertial measurement unit[15]. The temporal phase ϕ_{temporal} is defined here as the proportion of a flap cycle of a leading bird at which a following bird initiates a flap. Spatial

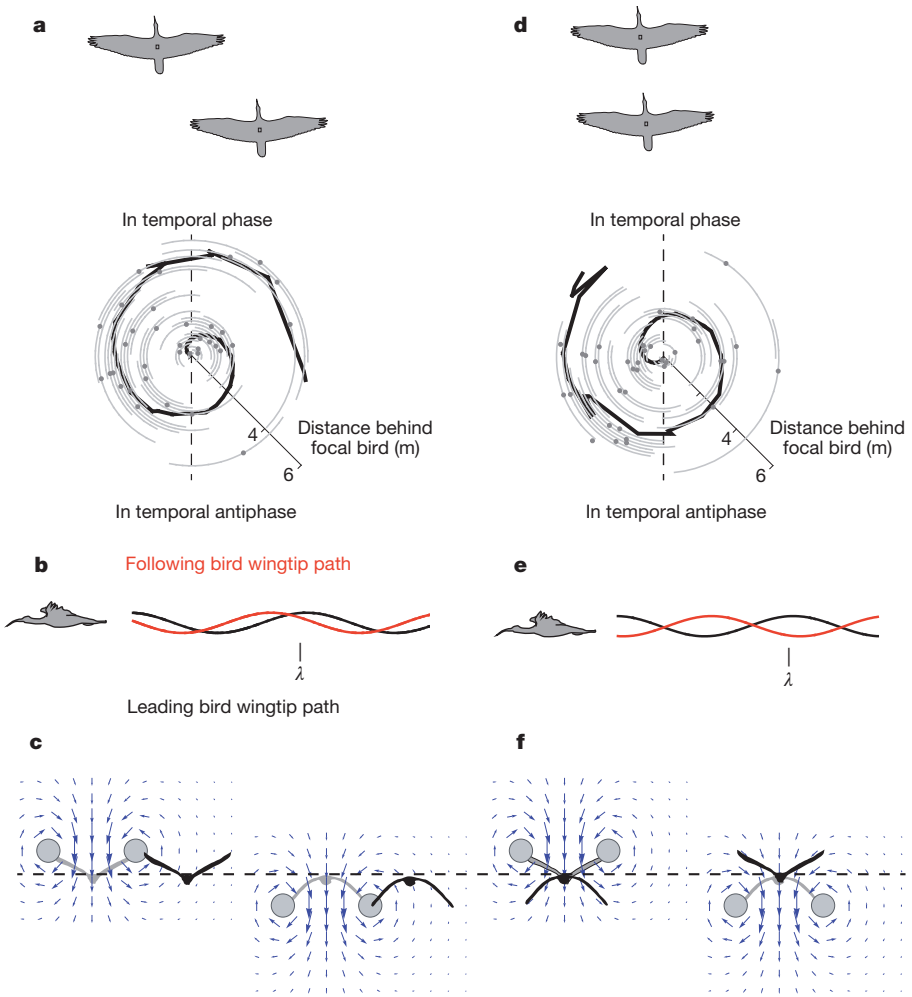


Figure 3 | Geometric and aerodynamic implications of observed spatial phase relationships for ibises flying in a V formation. Temporal phase increases as a function of position behind more advanced birds (median ± 95% confidence intervals of phase for each mean bird–bird interaction in a region). When positioned close to a wavelength in line with the V favoured position (**a**–**c**), wingtip paths approximately match: observed temporal phases agree with those predicted from the significant spatial phase relationship (thick black lines, ±95% confidence intervals) at the most populated 1 m × 1 m region, using the mean wavelength measured for each position. When positioned directly in line (**d**–**f**), following birds flap in spatial antiphase, maximally separating wingtip paths. In this case the model line is derived from the median spatial phase for all bird–bird interactions up to 4 m directly behind. Induced flow velocities (blue arrows, **c**, **f**), caused by the trailing wingtip vortices of the bird ahead (vortex cores denoted by grey circles), are modelled as infinitely long, parallel vortex filaments. Birds flying in typical V formation keep their wings close to the region of maximal induced upwash (**c**) throughout the flap cycle. Birds flying directly behind flap in spatial antiphase, potentially reducing the adverse effects of downwash (**f**), both in terms of magnitude and direction. For scale, the downwash directly between the vortices would be $(-)0.3\ \mathrm{m\ s^{-1}}$, between trailing vortices behind a bird of mass 1.3 kg, span 1.2 m at a speed of $15\ \mathrm{m\ s^{-1}}$ (no account is taken of flapping, viscosity or wake contraction). Alternative representations of **a** and **d** as Cartesian plots can be found in Supplementary Fig. 3, and Supplementary Fig. 4 details the extended data array shown beyond the presented model line.

phase ϕ_{spatial} makes use of the temporal phases, and takes account of the number of wavelengths, λ, between the bird ahead and the bird behind:

$$\phi_{\mathrm{spatial}} = \phi_{\mathrm{temporal}} - 2\pi\lambda.$$

A spatial phase of zero would indicate that, were the birds to be directly following each other, the wingtip paths would match.

In the most populated 1 m × 1 m favoured V position (Fig. 1c), Rayleigh's test[17] for circular statistics indicates a significant unimodal bias in both temporal (Rayleigh, $P = 0.018$, mean phase = 0.857; Hodges–Ajne's test, $P = 0.012$) and, more strongly, spatial (Rayleigh, $P = 0.003$, mean phase = −1.155; Hodges–Ajne, $P = 0.004$) phases (Fig. 3a, b) (see Supplementary Table 1 for further statistics; Supplementary Figs 2a, 3a and 4a). Flapping in spatial phase indicates that the wing of a following bird goes up and down tracking the path through the air previously described by the bird ahead. The following bird then benefits from consistently flapping into the upwash region from the preceding bird (Fig. 3b, c), presumably reducing the power requirements for weight support[12,14].

In contrast, birds flying directly behind, tracking the bird ahead in a streamwise position (sampled region 0.5 m across, 4 m streamwise, Fig. 1c), flap in close to spatial antiphase (median = 2.897, where precise antiphase would be ±3.142), significantly ($P < 0.05$) deviating from flapping 'in' spatial phase (see Supplementary Table 1 for further statistics; Supplementary Figs 2b, 3b and 4b). As such, the wingtip paths of the following bird do not match those of the preceding bird, and the wingtip paths are close to maximally separated. Birds flying directly behind another one in a streamwise location flap in spatial antiphase (Fig. 3d, e; see also Supplementary Figs 2b and 3b), potentially reducing the adverse effects of downwash (Fig. 3f), both in terms of magnitude and direction. If this position was aerodynamically adaptive, it would be predicted to be favoured at higher speeds, where parasite power is relatively high[18], compared with the induced power costs of weight support; forms of slipstreaming can reduce the drag experienced by followers[5,6,8,19], even in cases where there is zero net

horizontal momentum flux in the wake (that is, drag = thrust)—as in steady swimming—owing to temporal or local spatial[5,20,21] fluctuations from mean wake conditions. Whether the position immediately behind is accidental or intentional, and whether it offers any aerodynamic advantage or cost, is currently unclear. However, the wing-beat phasing observed when in this position would serve to displace the following bird's wings from regions of greatest downwash (presumably immediately inboard of the trailing wingtip vortices, close to wingtip paths described by the previous bird), through most of the flap cycle.

In transects both directly streamwise and along the favoured V position (Fig. 1c), temporal phase increases proportionally with distance behind the focal bird (Fig. 3a, d), with a full 2π cycle change in phase over a complete wavelength; spatial phase is approximately maintained up to 4 m behind the leading bird. Previously, there was much uncertainty about spatial wing-beat phasing and wingtip path coherence in flapping organisms. The only previous biological evidence of this phenomenon has come from tethered locusts, where distance manipulations between a leading locust and a follower altered the phase patterns of their wing-beats[22,23]. Physical models also support the potential for aerodynamic advantage due to phasing: appropriate timing between tandem flapping in model dragonfly wings improves aerodynamic efficiency[24]. Theoretical engineering models have taken into consideration flapping flight, and the extra benefits a flapping wing may accrue in formation flight[12,14]. Such models have suggested that upwards of 20% variation exists in the induced power savings to be gained, if flapping is done optimally in spatial phase, compared with out of phase[12] (Supplementary Fig. 4).

Here we have shown that ibis flight in V formation does, on average, match predictions of fixed-wing aerodynamics (Fig. 1c, d), but that flock structure is highly dynamic (Fig. 2). Further, temporal phasing of flapping relates both to streamwise and to spanwise position. This indicates remarkable awareness of, and ability to respond to, the wing-path—and thereby the spatial wake structure—of nearby flock-mates. Birds flying in V formation flap with wingtip path coherence—the wingtips take the same path—placing wings close to the oscillating positions of maximal upwash. In contrast, birds flying in line flap in spatial antiphase—the wingtip paths are maximally separated—consistent with avoidance of adverse downwash. This raises the possibility that, in contrast to conventional aircraft, following birds may be able to benefit from 'drafting' while, to a certain extent, avoiding an increased cost of weight support by evading localized regions of downwash. Optimal flight speeds would differ between solo flight, V formation flight and (whether net-beneficial or not) in-line flight, potentially providing some account for the unstable, dynamic nature of V formation flocks.

METHODS SUMMARY

Measurements. We equipped 14 juvenile northern bald ibises with back-mounted synchronized GPS (5 Hz) and inertial measurement units (300 Hz), mass 23 g (Supplementary Fig. 8), which were custom made in our laboratory, and tested and validated for accuracy and precision[15,16]. At the start of migration, the mass of the birds was 1.30 ± 0.73 kg, the 23 g loggers constituting approximately 3% of the body mass of the smallest bird. This is below the recommended 5% for flying animals[25]. The ibises formed part of a large-scale conservation programme. They had been hand-reared at Zoo Vienna (Austria), imprinted onto human foster parents and taught to follow a powered parachute (paraplane) to learn the migration routes (Methods). Experimental protocols were approved by the Royal Veterinary College local Ethics and Welfare Committee. A GPS trace of the ibis flight imposed over Google Earth (Landsat) can be found in Supplementary Data 1 as a KML file. GPS data were post-processed using GravNav Waypoint™ software[15,26], and inertial measurement unit data by custom-written MATLAB (R2012b, Mathworks) programs[16,26]. Mean flap frequency, speed and peak detection protocols are detailed in Supplementary Figs 5 and 6. For further details on post-processing, see Methods.

Statistical analysis. Circular statistics[17] were done in LabVIEW (National Instruments). First-order (Rayleigh test) and second-order (Hodges–Ajne) statistics were used to test the phasing of wing beats for significant deviations from random distribution. For further details on statistical analysis, see Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 May; accepted 3 December 2013.

1. Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. Effective leadership and decision-making in animal groups on the move. *Nature* **433,** 513–516 (2004).
2. Nagy, M., Akos, Z., Biro, D. & Vicsek, T. Hierarchical group dynamics in pigeon flocks. *Nature* **464,** 890–894 (2010).
3. May, R. M. Flight formations in geese and other birds. *Nature* **282,** 778–780 (1979).
4. Lissaman, P. B. & Schollenberger, C. A. Formation flight of birds. *Science* **168,** 1003–1005 (1970).
5. Liao, J. C., Beal, D. N., Lauder, G. V. & Triantafyllou, M. S. Fish exploiting vortices decrease muscle activity. *Science* **302,** 1566–1569 (2003).
6. Bill, R. G. & Hernnkind, W. F. Drag reduction by formation movement in spiny lobsters. *Science* **193,** 1146–1148 (1976).
7. Weimerskirch, H., Martin, J., Clerquin, Y., Alexandre, P. & Jiraskova, S. Energy saving in flight formation. *Nature* **413,** 697–698 (2001).
8. Fish, F. E. Kinematics of ducklings swimming in formation: consequence of position. *J. Exp. Zool.* **273,** 1–11 (1995).
9. Badgerow, J. P. & Hainsworth, F. R. Energy savings through formation flight? A re-examination of the vee formation. *J. Theor. Biol.* **93,** 41–52 (1981).
10. Cutts, C. J. & Speakman, J. R. Energy savings in formation flight of pink-footed geese. *J. Exp. Biol.* **189,** 251–261 (1994).
11. Hummel, D. Aerodynamic aspects of formation flight in birds. *J. Theor. Biol.* **104,** 321–347 (1983).
12. Willis, D. J., Peraire, J. & Breuer, K. S. in *Proc. 25th American Institute of Aeronautics and Astronautics Appl. Aerodynam. Conf.* http://doi.org/10.2514/6.2007-4182 (2007).
13. Hainsworth, F. R. Precision and dynamics of positioning by Canada geese flying in formation. *J. Exp. Biol.* **128,** 445–462 (1987).
14. Maeng, J. S. *et al.* A modelling approach to energy savings of flying Canada geese using computational fluid dynamics. *J. Theor. Biol.* **320,** 76–85 (2013).
15. Usherwood, J. R., Stavrou, M., Lowe, J. C., Roskilly, K. & Wilson, A. M. Flying in a flock comes at a cost in pigeons. *Nature* **474,** 494–497 (2011).
16. Wilson, A. M. *et al.* Locomotion dynamics of hunting in wild cheetahs. *Nature* **498,** 185–189 (2013).
17. Fisher, N. I. *Statistical Analysis of Circular Data* Ch. 4, 59–102 (Cambridge Univ. Press, 1993).
18. Pennycuick, C. J. *Bird Flight Performance: A Practical Calculation Manual* Ch. 3, 37–78 (Oxford Univ. Press, 1989).
19. Spence, A. J., Thurman, A. S., Maher, M. J. & Wilson, A. M. Speed, pacing and aerodynamic drafting in thoroughbred horse racing. *Biol. Lett.* **8,** 678–681 (2012).
20. Chatard, J.-C. & Wilson, B. Drafting distance in swimming. *Med. Sci. Sports Exerc.* **35,** 1176–1181 (2003).
21. Delextrat, A. *et al.* Drafting during swimming improves efficiency during subsequent cycling. *Med. Sci. Sports Exerc.* **35,** 1612–1619 (2003).
22. Kutsch, W., Camhi, J. & Sumbre, G. Close encounters among flying locusts produce wing-beat coupling. *J. Comp. Physiol. A* **174,** 643–649 (1994).
23. Camhi, J. M., Sumbre, G. & Wendler, G. Wing-beat coupling between flying locusts pairs: preferred phase and life enhancement. *J. Exp. Biol.* **198,** 1051–1063 (1995).
24. Usherwood, J. R. & Lehmann, F. Phasing of dragonfly wings can improve efficiency by removing swirl. *J. R. Soc. Interface* **5,** 1303–1307 (2008).
25. White, C. R. *et al.* Implantation reduces the negative effects of bio-logging on birds. *J. Exp. Biol.* **216,** 537–542 (2013).
26. King, A. J. *et al.* Selfish-herd behaviour of sheep under threat. *Curr. Biol.* **22,** R561–R562 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements The Waldrappteam assisted with data collection and provided logistical support (J.F., B.V.). We thank members of the Structure & Motion Laboratory for discussions and assistance, particularly J. Lowe, K. Roskilly, A. Spence and S. Amos, and C. White and R. Bomphrey for reading an earlier draft of the paper. Funding was provided by an Engineering and Physical Sciences Research Council grant to A.M.W., J.R.U. and S.Ha. (EP/H013016/1), a Biotechnology and Biological Sciences Research Council grant to A.M.W. (BB/J018007/1) and a Wellcome Trust Fellowship (095061/Z/10/Z) to J.R.U.

Author Contributions S.J.P., S.Ha., A.M.W. and J.R.U. developed the concept of the paper. J.F., S.He. and D.T. reared and trained the birds. S.J.P., S.He., D.T., B.V. and J.F. collected the field data. S.J.P., T.Y.H. and J.R.U. undertook the data processing and analyses; J.R.U. performed the circular statistics. S.J.P., T.Y.H, A.M.W. and J.R.U wrote the manuscript, with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.J.P. (SPortugal@rvc.ac.uk).

METHODS

Birds. Northern bald ibises (*G. eremita*) ($n = 14$, five females and nine males) were hatched at Zoo Vienna, Austria, in March 2011, and imprinted immediately onto human foster parents (S.He. and D.T.). At 4 months of age, the birds began training flights behind a powered parachute (paraplane). Training flights lasted between 1 and 4 h, and were up to 5 km in length. At the end of July, birds were fitted with dummy loggers to prepare them for being equipped with data loggers for the long-distance migratory flights. The mass of the birds at the start of migration was 1.30 ± 0.73 kg. As such, the 23 g loggers constituted approximately 3% of the body mass of the smallest bird. This is comfortably below the recommended 5% for flying animals[25,27]. Experimental protocols were approved by the Royal Veterinary College local Ethics and Welfare Committee The loggers were externally attached, using Velcro and a harness (Supplementary Fig. 8). The dummy loggers remained on when birds were at rest in the aviary, which was at all times apart from the migratory flights. The first migratory flight began in August. The total migratory flight plan was from the training site near Salzburg, Austria (47.75377° N, 13.052959° E), to Orbetello, Italy (42.425484° N, 11.232662° E). Once en route, birds were flown, on average, every third day. During flights, the birds followed the paraplane, but were typically to the side of the vehicle, on average 147 m laterally, consistently to the left, except for one turn (see Supplementary Figs 5 and 6). All loggers functioned fully. The birds were flown early in the morning (7:00 departure): later flight times increased the occurrence of thermalling and gliding, resulting in the birds not following the paraplane sufficiently. A GPS trace of the full flight, imposed over Google Earth (Landsat), can be seen in Supplementary Data 1 (as a KML file). The recorded flight was the second stage of the migration.

Data loggers. Further information about the loggers can be found in refs 15 and 16. Briefly, GPS was recorded at 5 Hz and data were post-processed differentially over the short baseline between base station and ibises, using Waypoint GrafNav 8.10. L1 coarse/acquisition (C/A) code pseudo-range measurements were used to calculate the position of each GPS logger, with velocity determined from L1 Doppler measurements. Using this approach can provide positional accuracy to 0.3 m and speed accuracy better than 0.1 m s^{-1}. Accelerometer data were recorded at 300 Hz.

Initial data processing. The flight was checked for any periods when the birds had maintained periods of circling flight (note we do include one circle in our sequence) through examination of the GPS and accelerometer traces, and these sections were removed (less than 4 min of the total flight duration). The remaining flight, therefore, consisted of straight-line flight. The take-off and landing periods were removed, as, when taking off, it took approximately 4 min for the birds to form a coherent flock, and to follow the paraplane. Similarly, when the paraplane began to descend at the end of the flight, the birds separated and began to glide during descent. The position of the paraplane was recorded and tracked by a data logger (see Supplementary Figs 5 and 6). The GPS, recorded at 5 Hz, was interpolated (MATLAB, R2012b, Mathworks) to the same sampling rate as the accelerometer data, at 300 Hz. The interpolation replaced missing values in the GPS. GPS and accelerometer data were passed through a fourth-order Butterworth filter (MATLAB). To produce the histograms (Fig. 1b, c), the original GPS values were used after being interpolated to a constant 5 Hz sampling frequency. In Figs 1b and 2, the colour scale refers to the duration a bird was present in each 0.25 m × 0.25 m grid. For Fig. 1c, the colour scale refers to the number of flaps recorded in each grid. In Fig. 1c, the regional transect labelled 'streamwise transect' is offset because, for display and analysis, all data from the left side are mirrored to the right so all data points are on one side; thus the centre of the first sampled region lies 0.125 m behind and to the right of the lead bird. Dorsal acceleration was used to determine each wing flap, and the upper reversal point[28] of the flap cycle (see Supplementary Figs 5 and 6). Note that this reversal point in acceleration of the back need not relate to peak wing elevation—or indeed any particular wing kinematic—for the phasing analysis to function.

Height. Height was recorded. The precision of height measurements, however, is lower than for horizontal positions[29]. This is because there were no satellites below the birds, and this geometry of the satellites caused a reduction in precision[29]. We do not consider vertical position because of the small 'signal' of interest (very slight vertical deflections) compared with the relatively high 'noise' (inevitable because of GPS satellite geometry). We chose a section where, according to the available error measurement calculated by Waypoint (see 'Data loggers'), the height values were relatively consistent and that during this flight portion the birds were flying close to the same horizontal plane.

Calculating flock formation and individual positioning. To establish positioning of individuals and structure in the flock, a flock centroid was determined. To calculate the centroid of the flock, the MATLAB function 'centroid' was used. This function calculates the centroid of a polygon. The MATLAB centroid function treated each bird as a point of a polygon, and determined the centroid for each time point. An average speed was calculated and any birds with a speed discrepancy higher than 3 m s^{-1} away from the mean flock speed were removed for that time point. From this, the resultant centroid was calculated now containing only birds close in position and speed. A rotation matrix was applied to the data to re-orient the heading so all birds were heading 'up', and the direction of the centroid was always in the positive 'Y' direction. The resultant matrix comprised a position for each bird for each sampling point. Theta (θ), the angle between each bird and the lead bird, was calculated, transforming Cartesian to polar coordinates (cart2pole, MATLAB). For data presentation in the histograms (Fig. 1b), the field of view was set to 15 m × 15 m, and the area was divided into a 60 × 60 grid of bins (0.25 m × 0.25 m). Position 0/0 is the centroid. The heat histograms are shown as contour plots with five contour levels.

During the 7 min section of V formation flight, individual birds showed a certain degree of positional infidelity in the V flock (Fig. 2; see also Supplementary Fig. 1 and Supplementary Video 1). Although individuals contributed to the statistical V formation, their positioning was inconsistent. Certain individuals showed general preferences for a particular area in the V formation, whether left, right, front or rear, but the variability in positioning resulted in no clear leader in the flock (Fig. 2). Navigational ability and kin selection have been proposed as principal drivers of leadership in V formation flight[30], with more experienced birds or parents of a family group taking the lead[30]. The ibis flock in the present study comprised birds of the same age (<1 yr old), with no previous navigational experience of the route and no parent–offspring relationships. The absence of immediate kin selection and learnt navigational ability as possible factors determining a V formation structure in the recorded flight strengthens the evidence for an aerodynamic function behind the V formation observed in the ibis. The young age of the birds, however, may be the main factor why there was a lack of a clear leader in the ibis flock, contrasting with previous observations of adult ibises, in which consistent leaders in flocks were identified[31]. Spontaneous and inconsistent leadership has been identified in bird flocks either where no consistent social hierarchy exists[32], or when no previous knowledge of a route is known[33]. For other 'classic' V formation fliers, the first migration is a significant cause of mortality for young birds, even when migrating with parents. As such, aerodynamic mechanisms that reduce the energetic cost of (albeit only very infrequent) migratory flight may present considerable selection advantage.

Movement in flock. Movement in the V formation was investigated by taking a 45° line, the preferred angle for positioning with the V (Fig. 1c) as a transect from the apex of the V. The apex was determined by the intersection of two 45° lines, down each side of the V formation. For every bird for each time point, we measured how far it was positioned from the 45° perpendicular transect line. For simplicity of analysis, all data were flipped (mirrored) so they could all be plotted against one 45° line. In Supplementary Fig. 1a, the red circles represent the original positions of the birds, for all birds and all times. From this, the shortest distance to the 45° line was calculated (blue line) and the position was projected on the 45° line; then the distance between projected position and green circle (the centroid) was calculated. The standard deviations are from the blue perpendicular line rather than the absolute distance, and represent how much the position varied with respect to the line. A mean (s.d.) was then calculated for each perpendicular/parallel relationship (Supplementary Fig. 1b, c). The positioning of all individuals varied more along the line than out from the line (Supplementary Fig. 1c). If the changes in position were due only to error in logger measurement, the variation in perpendicular and parallel distance and position would be expected to be equal. Because most of the variation is present along the line, the variation can be confidently attributed predominantly to bird movement, not logger noise.

Circular statistics and phasing analysis. Circular statistics were applied using LabVIEW (National Instruments), following that of Fisher[17,34–36].

The relative positions (in the direction of flight) and phase relationships (as a proportion of the flap cycle of each 'ahead' bird) were determined for every bird following another individual. Determining appropriate independent sample criteria when considering phases is vital[36], and presents a challenge when analysing phase relationships. Consider the case of two birds flying at the same relative position and at the same frequency; they would maintain the same phase relationship indefinitely. Each flap would certainly not be considered an independent sample. As a conservative alternative, we take a mean phase for any bird–bird pairing for a given area to be an independent sample; no account is taken of the length of time or number of flaps spent in the area. Perversely, this technique actually makes use of the variability in relative position, and would be poor for absolutely rigid V formations.

Statistical tests[17,34–36] analysed two regions, combining left and right sides: one representing V formation flight (from 0.49 to 1.49 m both spanwise and streamwise), containing the highest density of flaps; the other for nose-to-tail, streamwise flight, covering a volume 0.25 m spanwise from midline (so 0.5 m behind) and 4 m behind. This provided $n = 165$ and $n = 160$ bird–bird pairs for V formation and nose-to-tail regions, respectively.

The Rayleigh test was applied to determine the presence of a single unimodal direction in phase without preconceptions of any mean direction. This found a significant departure from randomness—a significant unimodal bias—in phase (whether temporal or spatial) for the V formation region. Both Rayleigh's test (parametric) and Hodges–Ajne's test[17,34–36] (non-parametric) on this region indicated that both the temporal and the spatial phases (taking into account the wavelength of whichever bird was ahead) were significantly different from those that would be found from a random distribution[37,38].

The median phase for a given region—and its 95% confidence intervals—allows a specified alternative to be tested against. Fig. 3a, d and Supplementary Fig. 3a, b show the median statistics in graphical form for the two regions. Zero or 'in' spatial phase falls outside the 95% confidence intervals for the nose-to-tail region.

The median spatial phases for the two regions described above were used to predict the temporal phases for 0.25 m × 0.25 m along two streamwise transects using the wavelength measured for each volume along the transect. If the median spatial phase was π—out of phase, as it is close to in the nose-to-tail transect—we would predict it to be π every integer number of wavelengths, and 0 or 'in' temporal phase at ½, 3/2, 5/2, etc. wavelengths. The model—with bounding confidence intervals due the spatial median—is shown as lines in Fig. 3a, d. Measured median temporal phases (±95% confidence intervals of the median) broadly match the predicted values (see also Supplementary Fig. 3a, b, which gives the same data in Cartesian form). Although the fit between model and observed temporal phases is visually convincing, formal statistical treatment is avoided because of uncertainty over independence between neighbouring spatial regions along the transects.

Modelled induced flow behind flapping birds. The implications of flap phasing in terms of potential interaction with induced flows are shown in Fig. 3c, f. For this model, it is assumed that the wingtip vortex left behind a bird ahead (the grey bird) of a follower (the black bird) follows the wingtip path through space—the convection of the vortex core (which, on average, will be inwards and downwards) is neglected[39,40]. Induced flow-fields are modelled following the Biot-Savart law[41,42], treating the wingtip vortices as infinitely long, parallel filaments; no account is taken of variation in lift throughout the wingstroke cycle. Induced flows near the vortex cores are not modelled; these regions are represented by grey circles. They, although being correct given the reductions and assumptions described, should not be taken as accurate quantitative calculations of the local flowfield. However, the principles they demonstrate—the strongest region of upwash and downwash close to outboard and inboard, respectively, of the wingtip path—meet basic aerodynamic expectations and recent modelling results[41,42]. For scale, the downwash directly between the vortices would be $(-)0.3\ \mathrm{m\,s^{-1}}$ between trailing vortices for behind a bird of mass 1.3 kg, span 1.2 m at a speed of $15\ \mathrm{m\,s^{-1}}$ (without modelling flapping or wake contraction).

27. Barron, D. G., Brawn, J. D. & Weatherhead, P. J. Meta-analysis of transmitter effects on avian behaviour and ecology. *Methods Ecol. Evol.* **1,** 180–187 (2010).
28. Norberg, U. M. *Vertebrate Flight: Mechanics, Physiology, Morphology, Ecology and Evolution* Ch. 9, 118–132 (Springer, 2011).
29. Kaplan, E. & Hegarty, C. *Understanding GPS: Principles and Applications* Ch. 7, 304–334 (Artech House, 2005).
30. Andersson, M. & Wallander, J. Kin selection and reciprocity in flight formation? *Behav. Ecol.* **15,** 158–162 (2003).
31. Petit, D. R. & Bildstein, K. L. Development of formation flying in juvenile white ibises (*Eudocimus albus*). *Auk* **103,** 244–246 (1986).
32. Rands, S. A., Cowlishaw, G., Pettifor, R. A., Rowcliffe, J. M. & Johnstone, R. A. Spontaneous emergence of leaders and followers in foraging pairs. *Nature* **423,** 432–434 (2003).
33. Biro, D., Sumpter, D. J. T., Meade, J. & Guilford, T. From compromise to leadership in pigeon homing. *Curr. Biol.* **16,** 2123–2128 (2006).
34. Mardia, K. & Jupp, P. *Directional Statistics* Ch. 6, 94–110 (Wiley, 1999).
35. Sprent, P. & Smeeton, N. C. *Applied Nonparametric Statistical Methods* Ch. 4, 83–122 (Taylor & Francis, 2007).
36. Batschelet, E. *Circular Statistics in Biology* Chs 9, 15 (Academic, 1981).
37. Wiltschko, W. *et al.* Lateralisation of magnetic compass orientation in a migratory bird. *Nature* **419,** 467–470 (2002).
38. Holland, R. A. *et al.* Testing the role of sensory systems in the migrating heading of a songbird. *J. Exp. Biol.* **212,** 4065–4071 (2009).
39. Hubel, T. Y. *et al.* Wake structure and wing kinematics: the flight of the lesser dog-faced fruit bat, *Cynopterus brachyotis*. *J. Exp. Biol.* **213,** 3427–3440 (2010).
40. Hubel, T. Y. *et al.* Changes in kinematics and aerodynamics over a range of speeds in *Tadarida brasiliensis*, the Brazilian free-tailed bat. *J. R. Soc. Interface* **9,** 1120–1130 (2012).
41. Kroner, E. Dislocations and the Biot-Savart law. *Proc. Phys. Soc. A* **68,** 53–55 (1955).
42. Griffiths, D. J. *Introduction to Electrodynamics* Ch. 5, 215 (Prentice Hall, 1998).

LETTER

doi:10.1038/nature12788

A mitochondrial genome sequence of a hominin from Sima de los Huesos

Matthias Meyer[1], Qiaomei Fu[1,2], Ayinuer Aximu-Petri[1], Isabelle Glocke[1], Birgit Nickel[1], Juan-Luis Arsuaga[3,4], Ignacio Martínez[3,5], Ana Gracia[3,5], José María Bermúdez de Castro[6], Eudald Carbonell[7,8] & Svante Pääbo[1]

Excavations of a complex of caves in the Sierra de Atapuerca in northern Spain have unearthed hominin fossils that range in age from the early Pleistocene to the Holocene[1]. One of these sites, the 'Sima de los Huesos' ('pit of bones'), has yielded the world's largest assemblage of Middle Pleistocene hominin fossils[2,3], consisting of at least 28 individuals[4] dated to over 300,000 years ago[5]. The skeletal remains share a number of morphological features with fossils classified as *Homo heidelbergensis* and also display distinct Neanderthal-derived traits[6–8]. Here we determine an almost complete mitochondrial genome sequence of a hominin from Sima de los Huesos and show that it is closely related to the lineage leading to mitochondrial genomes of Denisovans[9,10], an eastern Eurasian sister group to Neanderthals. Our results pave the way for DNA research on hominins from the Middle Pleistocene.

The Sima de los Huesos site (see Fig. 1 for a map) is located at the foot of a 13 m vertical shaft, about 30 m below the surface and 500 m from the closest current entrance to the karst system[11]. Humidity at the site is close to saturation, temperature in the cave is constant around 10.6 °C and the fossils have been protected from major disturbances since deposition[12]. The Sima de los Huesos is also noteworthy because it has provided unique evidence of long-term DNA survival. DNA preservation in the site was first proposed based on enzymatic amplification of a few short mitochondrial DNA (mtDNA) fragments from Middle Pleistocene cave bear remains[13]. Recently, improvements in DNA extraction[14] and library preparation[10] techniques for highly degraded ancient DNA have enabled the retrieval of a complete mitochondrial genome of a cave bear (*Ursus deningeri*) found with the hominin remains in the cave[14]. DNA preservation for hundreds of thousands of years has otherwise been documented only under permafrost conditions[15,16].

Figure 1 | Location of the Middle Pleistocene site of Sima de los Huesos (yellow) as well as Late Pleistocene sites that have yielded Neanderthal DNA (red) and Denisovan DNA (blue).

To investigate whether DNA may also be preserved in the hominin remains, we obtained several samples of bone, totalling 1.95 g, by drilling holes into the breaks of a femur (Femur XIII, ref. 17) excavated in three parts, one in 1994 and the other two in 1999 (Fig. 2). DNA was isolated using a recently published silica-based method[14] and converted into 77 libraries for sequencing[10,18] (Extended Data Table 1). Following library amplification, we first characterized a subset of the libraries by shallow shotgun sequencing on Illumina's MiSeq platform (Extended Data Fig. 1). Overlapping paired-end reads were merged to reconstruct full-length molecule sequences and mapped against the human genome using Burrows–Wheeler alignment (BWA)[19]. For most libraries, fewer than 0.1% of the sequences could be confidently aligned to the human genome (Extended Data Table 2), but 21 libraries yielded proportions of aligned sequences that were high enough (between 0.1% and 8.4%) to investigate the frequencies of C to T substitutions at sequence ends, which are increased in authentic ancient DNA due to accelerated cytosine deamination in single-stranded overhangs[20–22]. However, in no case did C to T substitution frequencies exceed 3% at 5′ ends and 6% at 3′ ends (Extended Data Table 2), indicating that those libraries that are rich in human DNA are dominated by present-day human contamination.

We next enriched all libraries for mtDNA, using a probe set based on a present-day human sequence. An initial inspection of the isolated sequences revealed the closest similarities to the mtDNA of a Denisovan, an extinct archaic group related to Neanderthals[9]. Therefore, the libraries were additionally enriched with probes based on the Denisovan mtDNA[23]. Sequencing was performed on Illumina's HiSeq 2500 platform from both ends, and overlapping reads were merged and aligned to the human reference mtDNA. Sequences with identical start and end coordinates, which often represent amplification products of the same starting molecules, were fused to create consensus sequences, and sequences shorter than 30 base pairs (bp) were discarded. The enriched libraries yielded a sufficient number of mitochondrial sequences to estimate the frequencies of C to T substitutions. These varied widely among the libraries, ranging from 1% to 45% at 5′ ends, and from 2% to 47% at 3′ ends (Extended Data Table 1). In agreement with the shotgun

Figure 2 | Femur XIII reassembled from three parts after sampling. The natural fractures are visible in the proximal third of the femur.

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [2]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China. [3]Centro de Investigación Sobre la Evolución y Comportamiento Humanos, Universidad Complutense de Madrid–Instituto de Salud Carlos III, 28029 Madrid, Spain. [4]Departamento de Paleontología, Facultad de Ciencias Geológicas, Universidad Complutense de Madrid, 28040 Madrid, Spain. [5]Área de Paleontología, Depto. de Geografía y Geología, Universidad de Alcalá, Alcalá de Henares, 28871 Madrid, Spain. [6]Centro Nacional de Investigación sobre la Evolución Humana, Paseo Sierra de Atapuerca, 09002 Burgos, Spain. [7]Institut Català de Paleoecologia Humana i Evolució Social, C/Marcel·lí Domingo s/n (Edifici W3), Campus Sescelades, 43007 Tarragona, Spain. [8]Àrea de Prehistòria, Dept. d'Història i Història de l'Art, Univ. Rovira i Virgili, Fac. de Lletres, Av. Catalunya, 35, 43002 Tarragona, Spain.

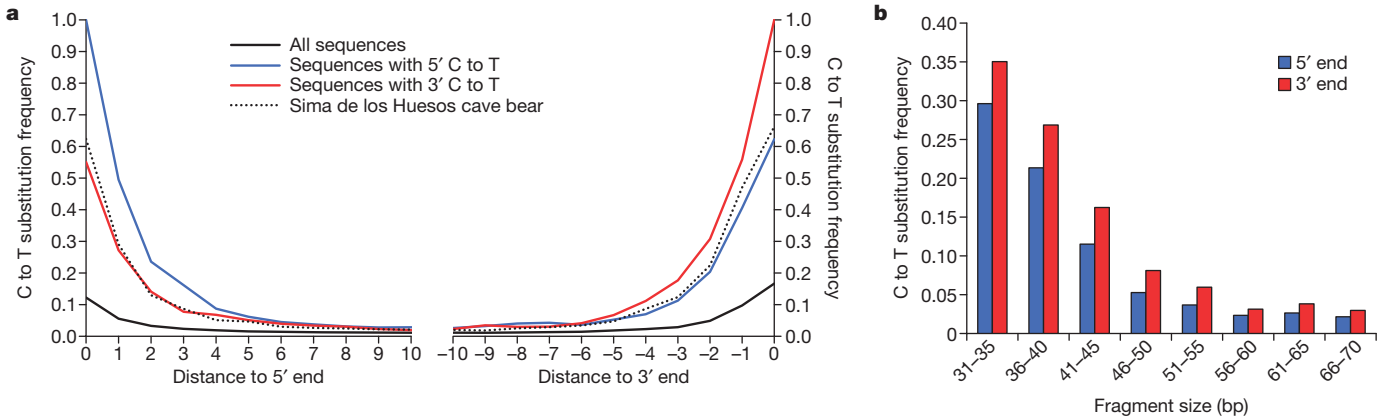


Figure 3 | Patterns of cytosine deamination in the libraries constructed from the Sima de los Huesos hominin femur. a, C to T substitution frequencies are shown for the terminal positions of the aligned sequences for all sequences (black), those sequences carrying a C to T substitutions at their 5′ ends (blue), at their 3′ ends (red), and for all Sima de los Huesos cave bear sequences from the *U. deningeri* sample[9] (dotted line). **b**, C to T substitution frequencies at the first and last base of sequences in different fragment length bins.

sequencing results, the libraries yielding the largest number of mitochondrial sequences exhibited very low terminal C to T substitution frequencies (≤3% and ≤6% at 5′ and 3′ ends, respectively; Extended Data Fig. 2) indicating that they are dominated by present-day human contamination. Libraries showing C to T substitution frequencies of less than 5% at either end were considered to be too contaminated and therefore disregarded in subsequent analyses.

Variation in C to T substitution frequencies among libraries suggest that two populations of sequences are present in the data, an endogenous population strongly affected by cytosine deamination and a contaminating population showing much less deamination. To test if this is the case, we determined the 5′ C to T substitution frequencies for sequences showing a 3′ C to T difference to the reference and vice versa, thereby enriching for putatively endogenous DNA. C to T substitution frequencies indeed increased to 55% at 5′ ends and 62% at 3′ ends, numbers that are close to those determined for the *U. deningeri* sample from Sima de los Huesos[14] (Fig. 3a). Furthermore, stratification of the deamination signal by fragment length shows that the endogenous DNA is primarily present among sequences that are shorter than 45 bp, again in agreement with the situation in the *U. deningeri* sample (Fig. 3b and Extended Data Fig. 3). Based on these results, we removed sequences longer than 45 bp and those that do not carry a terminal C to T substitution on either the 5′ or 3′ end (Extended Data Table 3). In addition, we applied a mapping quality filter to ensure unique placement of the sequences within the mtDNA genome and readjusted the alignment parameters to tolerate up to five C to T differences but no more than two other differences to the reference mtDNA sequence to discriminate against spurious alignments. Finally, T bases at the first and last three positions of each sequence were masked to reduce the impact of deamination-induced substitutions during consensus calling.

We first called consensus bases for 15,181 positions of the mitochondrial genome that were covered by 5 or more sequences of which at least 80% agreed. Average coverage across these positions was 21.8. However, such strict filtering increases the risk of ascertainment bias because residual modern human contamination as well as capture and mapping biases may lead to the exclusion of positions where the Sima de los Huesos specimen differs from the probes or the reference sequence. We therefore built a second more inclusive consensus by considering the three terminal positions while selecting sequences with C to T substitution and lowering the requirements for coverage and consensus agreement to 3 and >67%, respectively. This consensus encompasses 16,302 positions or ~98% of the human mitochondrial reference genome, with an average coverage of 31.6 (Extended Data Fig. 4). Third, to evaluate whether the use of Denisovan capture probes influence the results, we built a consensus using the strictest filtering criteria described above, but including only sequences isolated with present-day human mtDNA probes (Extended Data Fig. 5).

We reconstructed phylogenetic trees in a Bayesian statistical framework[24] using the three Sima de los Huesos consensus mtDNA sequences as well as the mtDNAs of present-day and ancient humans, Neanderthals, Denisovans, chimpanzees and bonobos. All three trees support a topology in which the Sima de los Huesos mtDNA shares a common ancestor with Denisovan mtDNAs to the exclusion of the other mtDNAs analysed with maximum posterior probability (Fig. 4 and Extended Data Fig. 6). As expected owing to its age, the branch leading to the Sima de los Huesos mtDNA is shorter than those leading to any of the other archaic or present-day humans. Using 13 directly dated ancient mtDNA sequences for calibration[25] and the three consensus sequences, we estimated the age of the Sima de los Huesos specimen based on the length of its mtDNA branch (Table 1). These dates vary between 0.15 to 0.64 million years with point estimates close to 400,000 years. This is in striking agreement with the point estimate of 409,000 years for the *U. deningeri* mtDNA[14]. We similarly estimated the divergence times of the major mitochondrial lineages (Table 1 and Extended Data Table 4) and find that the estimates for the divergence of the mtDNAs of the

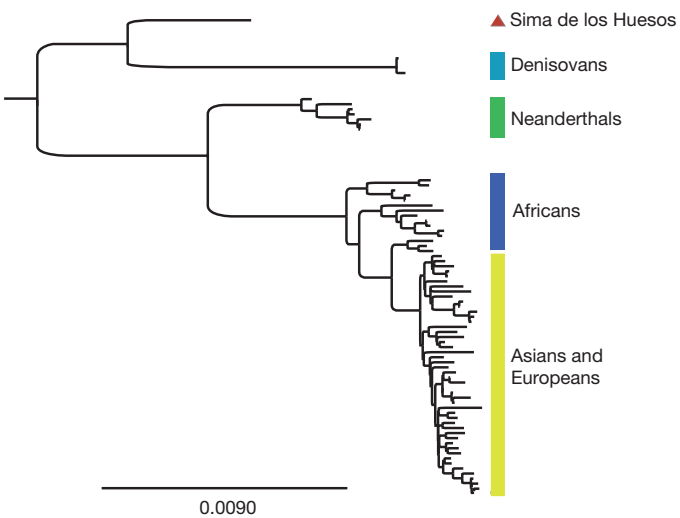


Figure 4 | Bayesian phylogenetic tree of hominin mitochondrial relationships based on the Sima de los Huesos mtDNA sequence determined using the inclusive filtering criteria. All nodes connecting the denoted hominin groups are supported with posterior probability of 1. The tree was rooted using chimpanzee and bonobo mtDNA genomes. The scale bar denotes substitutions per site.

Table 1 | Divergence times of the major hominin mtDNA lineages and the age of the Sima de los Huesos specimen as estimated by using three different filtering strategies for consensus calling

Mitochondrial lineages	Divergence dates and molecular age estimates in Myr BP (95% highest posterior density (HPD) intervals in brackets)		
	Strict filters	Inclusive filters	Strict filters enriched with human probes only
Sima de los Huesos age estimate	0.38 (0.18–0.60)	0.37 (0.16–0.59)	0.40 (0.15–0.64)
Human – Neanderthal	0.51 (0.35–0.69)	0.53 (0.36–0.72)	0.44 (0.29–0.61)
Sima de los Huesos/Denisova – Human/Neanderthal	0.99 (0.70–1.33)	1.04 (0.72–1.41)	0.83 (0.54–1.11)
Sima de los Huesos – Denisova	0.73 (0.50–1.01)	0.76 (0.51–1.06)	0.65 (0.40–0.92)

Sima de los Huesos hominin and Densiovans vary between 0.40 and 1.06 million years with point estimates around 700,000 years ago.

The fact that the Sima de los Huesos mtDNA shares a common ancestor with Denisovan rather than Neanderthal mtDNAs is unexpected in light of the fact that the Sima de los Huesos fossils carry Neanderthal-derived features (for example, in their dental, mandibular, midfacial, supraorbital and occipital morphology[2,6,7,26]). Denisovans were identified in 2010 based on DNA sequences retrieved from a manual phalanx and a molar found in southern Siberia[9,23]. Based on analyses of their nuclear genome[9,10] they are a sister group of Neanderthals, although the mtDNAs of Neanderthals and present-day humans share an mtDNA ancestor more recently with each other than with Denisovans[23]. This may be owing to either incomplete lineage sorting in the common ancestral populations of these groups or to gene flow into Denisovans from another archaic group[9].

Several evolutionary scenarios are compatible with the presence of a mtDNA sequence that falls on the Denisovan mtDNA lineage in a ~400,000-year-old hominin in western Europe. First, the Sima de los Huesos hominins may be closely related to the ancestors of Denisovans. However, this seems unlikely, because the presence of Denisovans in western Europe would indicate an extensive spatial overlap with Neanderthal ancestors, raising the question how the two groups could genetically diverge while overlapping in range. Furthermore, although almost no morphological information is available for Denisovans, a molar that carries Denisovan DNA is of exceptionally large size[9] and does not exhibit the cusp reduction seen in the Sima de los Huesos hominins[7]. Most importantly, the Sima de los Huesos specimen is so old that it probably predates the population split time between Denisovans and Neanderthals, which is estimated to one-half to two-thirds of the time to the split between Neanderthals and modern humans, which is estimated to be 170,000 to 700,000 years ago[9]. Second, it is possible that the Sima de los Huesos hominins represent a group distinct from both Neanderthals and Denisovans that later perhaps contributed the mtDNA to Denisovans. However, this scenario would imply the independent emergence of several Neanderthal-like morphological features in a group unrelated to Neanderthals. Third, the Sima de los Huesos hominins may be related to the population ancestral to both Neanderthals and Denisovans. Considering the age of the Sima de los Huesos remains and their incipient Neanderthal-like morphology, this scenario seems plausible to us, but it requires an explanation for the presence of two deeply divergent mtDNA lineages in the same archaic group, one that later recurred in Denisovans and one that became fixed in Neanderthals, respectively. A forth possible scenario is that gene flow from another hominin population brought the Denisova-like mtDNA into the Sima de los Huesos population or its ancestors. Such a hominin group might have also contributed mtDNA to the Densiovans in Asia[9,10]. Based on the fossil record, more than one evolutionary lineage may have existed in Europe during the Middle Pleistocene[27]. Several fossils have been found in Europe as well as in Africa and Asia that are close in time to Sima de los Huesos but do not exhibit clear Neanderthal traits. These fossils are often grouped into *H. heidelbergensis,* a taxon that is difficult to define[8,28,29], particularly with regard to whether the Sima de los Huesos hominins should be included[8]. Furthermore, there may have been relict populations of still earlier hominins, notably those classified as *Homo antecessor,* which share some morphological traits with Asian *Homo erectus*[30] and have been found just a few hundred metres away from Sima de los Huesos in Gran Dolina.

Although nuclear sequence data are needed to clarify the genetic relationship of the Sima de los Huesos hominins to Neanderthals and Densiovans, the mtDNA sequence establishes an unexpected link between Denisovans and the western European Middle Pleistocene fossil record. Future efforts will now focus on describing the mtDNA variation of the Sima de los Huesos hominins and retrieving nuclear DNA sequences from them. The latter will be a huge challenge given that almost two grams of bone were required to generate the mtDNA sequence even though several hundred copies of mtDNA exist per cell. Although preservation of DNA for such long periods of time may be favoured by unique preservation conditions in the Sima de los Huesos, the present results show that ancient DNA sequencing techniques have become sensitive enough to warrant further investigation of DNA survival at sites where Middle Pleistocene hominins are found.

METHODS SUMMARY

A detailed description of the methods used for data generation and analysis is provided in the Methods section.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 September; accepted 17 October 2013.
Published online 4 December 2013.

1. Carbonell, E. *et al.* The first hominin of Europe. *Nature* **452,** 465–469 (2008).
2. Arsuaga, J. L., Martinez, I., Gracia, A. & Lorenzo, C. The Sima de los Huesos crania (Sierra de Atapuerca, Spain). A comparative study. *J. Hum. Evol.* **33,** 219–281 (1997).
3. Arsuaga, J. L. *et al.* Size variation in Middle Pleistocene humans. *Science* **277,** 1086–1088 (1997).
4. Bermúdez de Castro, J. M. & Nicolas, M. E. Palaeodemography of the Atapuerca-SH Middle Pleistocene hominid sample. *J. Hum. Evol.* **33,** 333–355 (1997).
5. Bischoff, J. L. *et al.* Geology and preliminary dating of the hominid-bearing sedimentary fill of the Sima de los Huesos Chamber, Cueva Mayor of the Sierra de Atapuerca, Burgos, Spain. *J. Hum. Evol.* **33,** 129–154 (1997).
6. Martínez, I. & Arsuaga, J. L. The temporal bones from Sima de los Huesos Middle Pleistocene site (Sierra de Atapuerca, Spain). A phylogenetic approach. *J. Hum. Evol.* **33,** 283–318 (1997).
7. Martinón-Torres, M., Bermudez de Castro, J. M., Gomez-Robles, A., Prado-Simon, L. & Arsuaga, J. L. Morphological description and comparison of the dental remains from Atapuerca-Sima de los Huesos site (Spain). *J. Hum. Evol.* **62,** 7–58 (2012).
8. Stringer, C. The status of *Homo heidelbergensis* (Schoetensack 1908). *Evol. Anthropol.* **21,** 101–107 (2012).
9. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
10. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
11. Ortega, A. I. *et al.* Evolution of multilevel caves in the Sierra de Atapuerca (Burgos, Spain) and its relation to human occupation. *Geomorphology* **196,** 122–137 (2013).
12. Arsuaga, J. L. *et al.* Sima de los Huesos (Sierra de Atapuerca, Spain). The site. *J. Hum. Evol.* **33,** 109–127 (1997).
13. Valdiosera, C. *et al.* Typing single polymorphic nucleotides in mitochondrial DNA as a way to access Middle Pleistocene DNA. *Biol. Lett.* **2,** 601–603 (2006).
14. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110,** 15758–15763 (2013).
15. Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317,** 111–114 (2007).
16. Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 (2013).

17. Carretero, J. M. *et al.* Stature estimation from complete long bones in the Middle Pleistocene humans from the Sima de los Huesos, Sierra de Atapuerca (Spain). *J. Hum. Evol.* **62,** 242–255 (2012).
18. Gansauge, M. T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols* **8,** 737–748 (2013).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
20. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104,** 14616–14621 (2007).
21. Krause, J. *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20,** 231–236 (2010).
22. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Paabo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7,** e34131 (2012).
23. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464,** 894–897 (2010).
24. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
25. Shapiro, B. *et al.* A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* **28,** 879–887 (2011).
26. Arsuaga, J. L., Martinez, I., Gracia, A., Carretero, J. M. & Carbonell, E. Three new human skulls from the Sima de los Huesos Middle Pleistocene site in Sierra de Atapuerca, Spain. *Nature* **362,** 534–537 (1993).
27. Arsuaga, J. L. Colloquium paper: terrestrial apes and phylogenetic trees. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 2), 8910–8917 (2010).
28. Hublin, J. J. Out of Africa: Modern human origins special feature: The origin of Neandertals. *Proc. Natl Acad. Sci. USA* **106,** 16022–16027 (2009).
29. Mounier, A., Marchal, F. & Condemi, S. Is *Homo heidelbergensis* a distinct species? New insight on the Mauer mandible. *J. Hum. Evol.* **56,** 219–246 (2009).
30. Carbonell, E. *et al.* An Early Pleistocene hominin mandible from Atapuerca-TD6, Spain. *Proc. Natl Acad. Sci. USA* **102,** 5674–5678 (2005).

Acknowledgements We thank J. Dabney, M. Dannemann, C. de Filippo, S. Lippold, K. Prüfer, M. Slatkin, M. Stiller, C. Valdiosera and B. Viola for discussions and comments on the manuscript; G. Renaud and U. Stenzel for help with sequence data processing; B. Höber and A. Weihmann for performing the sequencing runs; M. Gansauge, P. Korlević, R. Rodríguez and I. Ureña for help in the laboratory; M. Schreiber for help with graphics; J. Trueba for providing the fossil image; M. Cruz Ortega for restoration of the fossil and the rest of the members of the Sima de los Huesos excavation team ?for decades of continuous efforts. Genetics work was funded by the Max Planck ?Society and its Presidential Innovation Fund. Field work at the Sierra de Atapuerca sites is funded by the Junta de Castilla y León and the Fundación Atapuerca. ?Research was supported by Spanish Ministerio de Ciencia e Innovación (project CGL2009-12703-C03) and Spanish Ministerio de Economía y Competitividad (project CGL2012-38434-C03).

Author Contributions M.M. designed the experiments and analysed the data; Q.F. performed phylogenetic analyses; A.A., I.G. and B.N. performed the experiments; J.-L.A., I.M., A.G., J.M.B. and E.C. excavated the fossil and provided expert archaeological and anthropological information; J.-L.A. and S.P. were involved in study design; and M.M., J.-L.A. and S.P. wrote the manuscript.

Author Information The Sima de los Huesos mtDNA consensus sequence (based on the inclusive filtering criteria) is deposited in GenBank under accession number KF683087. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (mmeyer@eva.mpg.de).

LETTER

doi:10.1038/nature12821

Perturbed neural activity disrupts cerebral angiogenesis during a postnatal critical period

Christina Whiteus[1,2], Catarina Freitas[1] & Jaime Grutzendler[1,2]

During the neonatal period, activity-dependent neural-circuit remodelling coincides with growth and refinement of the cerebral microvasculature[1,2]. Whether neural activity also influences the patterning of the vascular bed is not known. Here we show in neonatal mice, that neither reduction of sensory input through whisker trimming nor moderately increased activity by environmental enrichment affects cortical microvascular development. Unexpectedly, chronic stimulation by repetitive sounds, whisker deflection or motor activity led to a near arrest of angiogenesis in barrel, auditory and motor cortices, respectively. Chemically induced seizures also caused robust reductions in microvascular density. However, altering neural activity in adult mice did not affect the vasculature. Histological analysis and time-lapse *in vivo* two-photon microscopy revealed that hyperactivity did not lead to cell death or pruning of existing vessels but rather to reduced endothelial proliferation and vessel sprouting. This anti-angiogenic effect was prevented by administration of the nitric oxide synthase (NOS) inhibitor L-NAME and in mice with neuronal and inducible NOS deficiency, suggesting that excessive nitric oxide released from hyperactive interneurons and glia inhibited vessel growth. Vascular deficits persisted long after cessation of hyperstimulation, providing evidence for a critical period after which proper microvascular patterning cannot be re-established. Reduced microvascular density diminished the ability of the brain to compensate for hypoxic challenges, leading to dendritic spine loss in regions distant from capillaries. Therefore, excessive sensorimotor stimulation and repetitive neural activation during early childhood may cause lifelong deficits in microvascular reserve, which could have important consequences for brain development, function and pathology.

The development of a cerebral microvascular network that precisely matches regional metabolic demands is crucial given the brain's high energy consumption and susceptibility to ischaemia[3]. Although major cerebral vessels form during embryonic development, microvascular sprouting and pruning continue into the neonatal stages[1], concurrent with synaptogenesis, axonal growth and gliogenesis. Common molecular pathways regulate angiogenesis and axonal growth[4], suggesting that coordinated mechanisms establish a microvascular network that meets the requirements of adjacent neural tissue. Although some studies suggest there is a link between neural activity and microvascular plasticity[5–9], this remains controversial and it is unclear whether neural activity regulates vascular development or if angiogenesis follows an autonomous developmental program[10].

To address this question, we examined the effects of neural activity on cerebral microvascular development in neonatal mice. First, we reduced sensory input to the barrel cortex by bilateral whisker trimming for 10 days beginning at postnatal day 15 (P15). This reduces spiking activity and metabolism[11], and affects dendritic spine dynamics[12] in the barrel cortex. We quantified vascular branch points and total length from confocal images of various vascular markers (Supplementary Fig. 1a–e and Supplementary Video 1) and found that this manipulation did not affect vascular density in the barrel cortex (Fig. 1a and Supplementary Fig. 2a, b). Moderate whisker stimulation by environmental enrichment over 10 days also had no effect on microvascular density (Fig. 1a and Supplementary Fig. 2a, c). Therefore baseline sensory activity does not modulate neonatal cortical angiogenesis.

Unexpectedly, more persistent and repetitive activity led to reduced vascular density. Exposure to diverse tones, natural sounds and white noise over 10 h daily from P15 to P25 caused robust reductions in vessel branching and length (Fig. 1b, c and Supplementary Fig. 2a), which increased in magnitude when stimulation was extended (Supplementary Fig. 2d). This effect was specific to the stimulated region, as vascular density was reduced in the primary auditory cortex but not in the cingulate cortex (Fig. 1b). We then tested the effect of sustained whisker stimulation by performing unilateral whisker trimming and exposing mice to continuous air current. Daily 10-h stimulation for 8 days led to significant reductions in microvascular density of the barrel cortex corresponding to the stimulated whiskers (Fig. 1d and Supplementary Fig. 2a). Similarly, 3 h of daily treadmill running for 5 days reduced vessel density specifically in the motor cortex (Fig. 1e and Supplementary Fig. 2a). Interestingly, vascular reductions following auditory or whisker stimulations were most apparent in cortical layers 2/3 and 4, whereas motor hyperactivity had a more significant effect in layers 5 and 6 (Supplementary Fig. 2e–g). This is likely to be due to the fact that sensory cortical layers 2/3 and 4 connect to afferent inputs from the thalamus that are most reliably activated following stimulation[13], whereas layer 5 efferent neurons in motor cortex are robustly activated during motor output[14].

Over a 10-day period between P15 and P25, repetitive stimulation led to differences of up to 13% in branching and 8% in length between control and stimulated mice. This slight discrepancy is not surprising as most vessels formed postnatally are short capillaries that contribute substantially to branching but less so to length[1]. Although this decrease may seem modest, it represents a 70% reduction in the numbers of new vascular branches formed and a 80% reduction in length growth when compared to un-manipulated animals (Fig. 1f, g and Supplementary Fig. 2a). In contrast to neonates, chronic auditory stimulation of adult mice did not affect vascular density (Fig. 1b). This is likely to be due to the fact that brain endothelial proliferation and sprouting become very restricted after the postnatal period[1].

To test whether epileptiform activity also affects neonatal angiogenesis, we administered the cholinergic muscarinic agonist pilocarpine for 10 days, which induced mild generalized seizures without tonic–clonic convulsions lasting for approximately 2 h (Supplementary Fig. 3). These seizures caused robust reductions of cortical vessel density in neonates (Fig. 1h and Supplementary Figs 2a and 4), but did not affect the adult microvasculature (Fig. 1h). We tested a second seizure mechanism, independent of cholinergic stimulation by intracortical injection of tetanus toxin, which blocks neurotransmission in inhibitory interneurons. This treatment also induced non-convulsive seizures, after which we observed significant vessel reductions in the cortex, contralateral (Fig. 1i) and ipsilateral (data not shown) to the injection. Botulinum toxin, a structurally similar molecule that blocks excitatory

[1]Department of Neurology, Yale University School of Medicine, New Haven, Connecticut 06511, USA. [2]Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA.

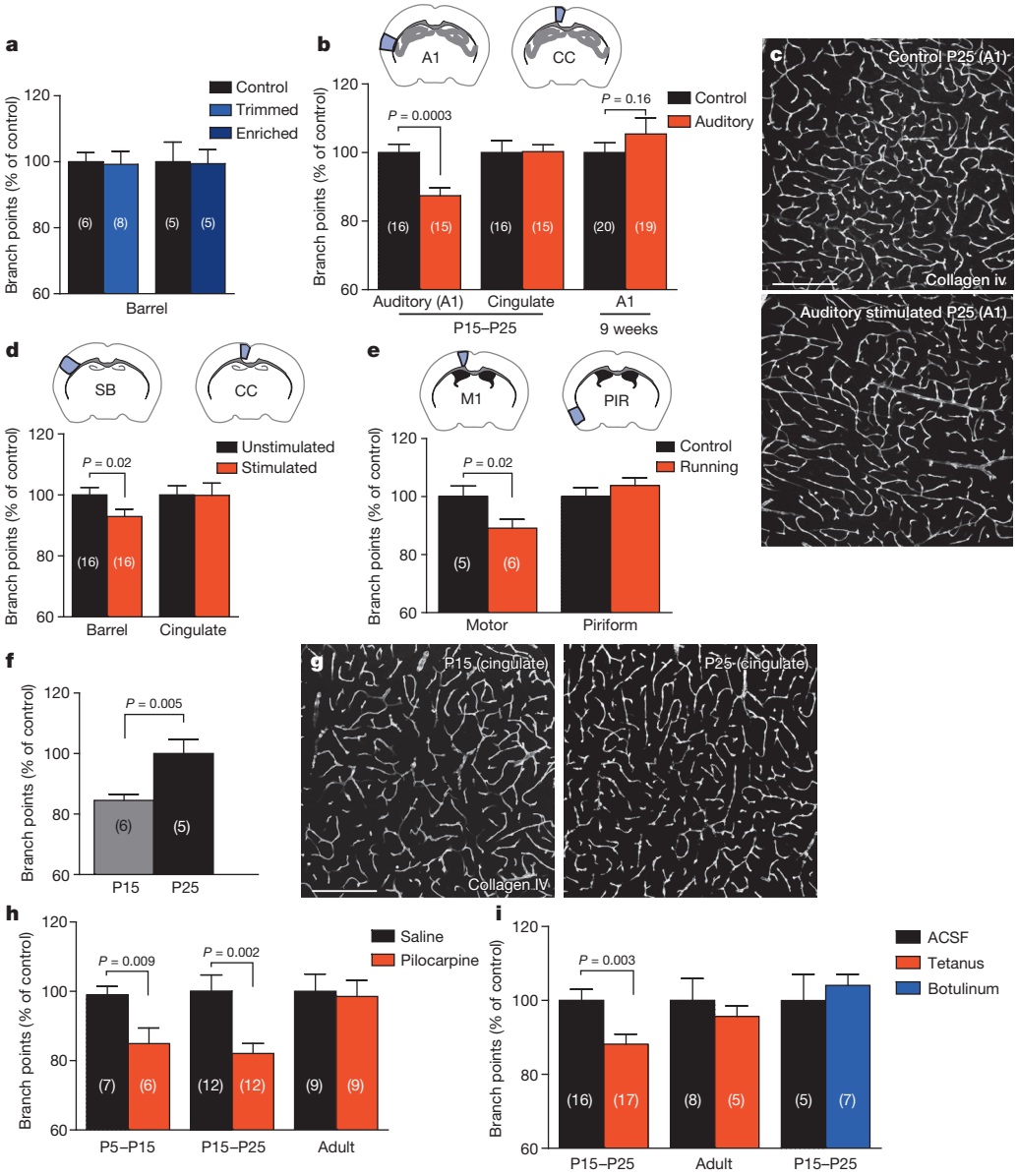


Figure 1 | Increased levels of neural activity during postnatal development lead to reduced microvascular density. **a**, Cortical microvascular density is not affected by reduced neural activity caused by whisker trimming or moderate enhancement of activity by environmental enrichment. **b–e**, Prolonged and repetitive activity through auditory stimulation using a variety of tones and sounds (**b**, **c**), increased unilateral deflection of whiskers by continuous air flow (**d**; stimulated hemisphere compared to unstimulated hemisphere contralateral to whisker-trimmed side), and running on a treadmill (**e**), cause reduced vessel branching in auditory (A1), sensory barrel (SB), and motor (M1) cortices respectively. Vessel density in control cortical areas (cingulate cortex, CC, and piriform, PIR) was unaffected. Auditory stimulation did not affect adult vasculature (**b**). **f**, **g**, Baseline cortical angiogenesis is robust between P15 and P25, as seen in representative images (**g**). **h**, **i**, Seizures caused by pilocarpine (**h**) or tetanus toxin (**i**) intracortical injections arrested vessel growth in neonates but not adults. Intracortical botulinum toxin injections caused no vessel changes (**i**). Scale bars, 200 μm (**c**, **g**). *n* numbers are given in brackets. *P* values, one-tailed student's *t*-test. Bars represent s.e.m.

neurotransmission, had no effect on vessels, supporting the conclusion that only hyperactivity affects microvascular remodelling (Fig. 1i).

To determine whether hyperactivity blocks vessel formation through stress or energetic depletion, we administered the glucocorticoid dexamethasone, a model of stress, or the glucose analogue, 2-deoxyglucose, to impair glucose metabolism. However, neither of these treatments affected the cortical vasculature (Fig. 2a). Furthermore, hyperstimulation was not associated with cell death, changes in cell density, or microglia activation (Supplementary Fig. 5). This suggests that reductions in vessel density are not secondary to homeostatic alterations, but a direct vascular response to hyperactivity.

We further examined the effect of hyperactivity on endothelial proliferation by repeated bromodeoxyuridine (BrdU) administration and found that both auditory stimulation and seizures markedly decreased proliferation selectively in endothelial cells (Fig. 2b–f and Supplementary Fig. 6). To determine how activity influenced vascular plasticity, we performed *in vivo* transcranial two-photon time-lapse imaging of mice expressing green fluorescent protein in their endothelium (Tie2-GFP). We found that administration of pilocarpine for 5 days caused 90% fewer microvascular formations (Fig. 2g, h, j), resulting in a significant reduction in vascular length added (Fig. 2k). In contrast, the number of eliminated branches was not different from controls (Fig. 2g, i, j). These results demonstrate that perturbed activity does not cause vascular regression, but instead arrests the proliferation and sprouting of new vessels.

Nitric oxide (NO) is released during neural activation and is known to modulate angiogenesis[15]. We therefore asked whether NO is involved

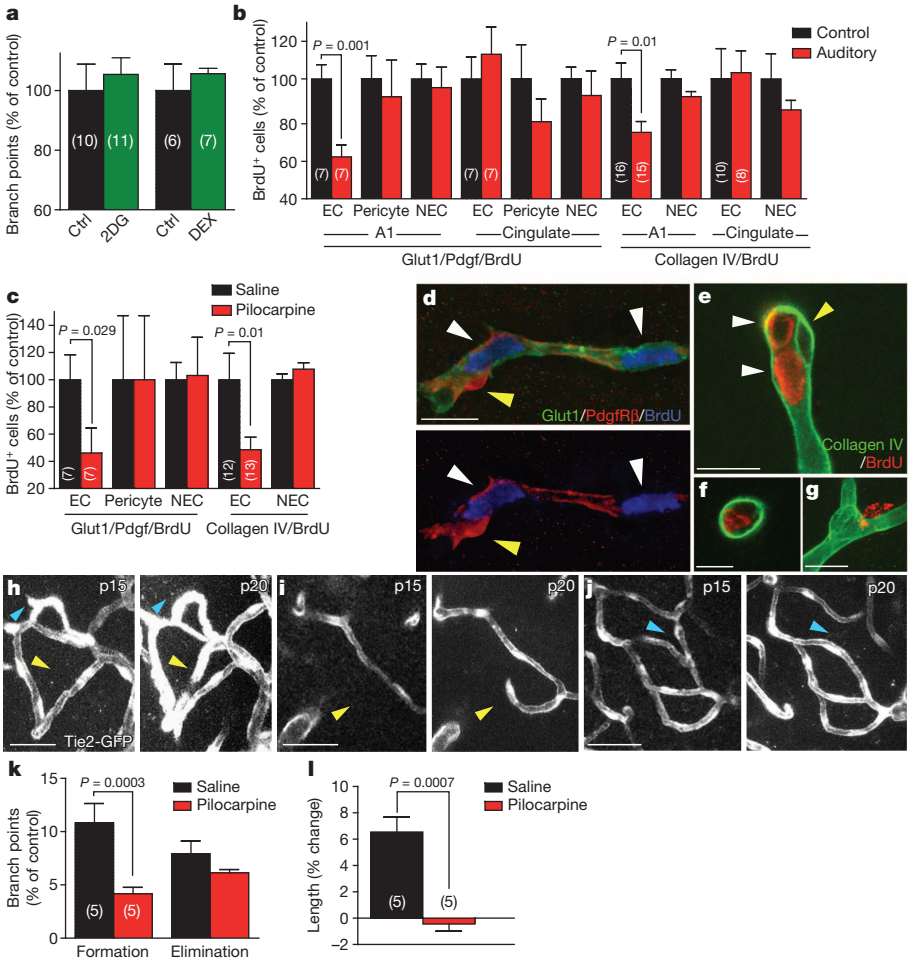


Figure 2 | Neural hyperactivity reduces endothelial proliferation and new vessel formations in the neonatal cortex. **a**, Daily treatment with 2-deoxyglucose (2DG) or dexamethasone (DEX) does not affect vessel density. **b**, Auditory stimulation reduces endothelial cell (EC) proliferation, measured by BrdU, in primary auditory but not in cingulate cortex control area. Non-endothelial cell (NEC) proliferation is not affected by auditory treatment in either location (see Supplementary Fig. 6). **c**, Pilocarpine injection cause decreased EC proliferation compared to saline-injected controls. NEC proliferation is not affected. **d**, Representative confocal image of cortical section stained with Glut1 (endothelium, green), Pdgf receptor-β (Pdgfrβ; pericytes, red), and BrdU (nuclei, blue) shows dividing endothelial cells (white arrowheads) near a non-proliferating pericyte (yellow arrowhead). **e**, Collagen IV and BrdU double staining shows two proliferating EC nuclei (white arrowheads; surrounded on one side by basal lamina) and a non-proliferating pericyte (surrounded luminally and abluminally by basal lamina; yellow arrowhead). **f**, Orthogonal views of an endothelial cell within the collagen IV basal lamina. **g**, Image of a proliferating pericyte (note that proliferating pericytes are not surrounded on all sides by detectable basal lamina). **h–l**, *In vivo* two-photon imaging reveals that hyperactivity inhibits new vessel formations. Representative two-photon time-lapse images show eliminations (blue arrowheads) of a sprout (**h**) and a vessel (**j**), and formations (yellow arrowheads) of a sprout (**i**) and a vessel (**h**) between time point 1 (left panel, P15) and 2 (right panel, P20). **k**, Percentage of vascular branches formed or eliminated during a 5-day treatment with either pilocarpine or saline. **l**, Significantly less vessel-length increase in pilocarpine-injected mice between P15 and P20 than in saline-injected controls. Scale bars, 10 μm (**d**, **e**, **g**), 5 μm (**f**), 50 μm (**h–j**). *P* values, one-tailed student's *t*-test. Bars represent s.e.m.

in the anti-angiogenic effects of neural activity. In the brain, NO is produced by three NOS isoforms: endothelial NOS (eNOS) and neuronal NOS (nNOS), which are constitutively active in endothelial cells and neurons, respectively; and inducible NOS (iNOS), which increases during inflammation, but is also present at baseline in glia and neurons[16–18]. Studies using NO donors or NOS overexpression have shown that NO is pro-angiogenic at moderate levels and anti-angiogenic at high physiological concentrations[19–21]. Notably, we found that activity-mediated vessel reductions were completely prevented by administration of a broad NOS inhibitor (L-NAME; L-NG-nitroarginine methyl ester), but not its enantiomer D-NAME (Fig. 3a, b). We then exposed isoform-specific loss-of-function mutants to auditory stimulation and found that mice lacking nNOS or iNOS were completely protected against activity-mediated vessel reductions, whereas eNOS-knockout mice were not (Fig. 3c, d). nNOS is expressed in subtypes of inhibitory interneurons[22], which are over-activated during prolonged stimulation, and glial iNOS also responds to changes in neuronal activity[17]. Therefore it is likely that activity-dependent NO production is responsible for the vascular effects observed (Supplementary Fig. 7 and Supplementary discussion 2). Although we found no changes in nNOS and iNOS protein levels following stimulation (Supplementary Fig. 8a), these isoforms produce NO very efficiently, therefore modest changes in their catalytic activity have strong effects on NO production. The anti-angiogenic effects of hyperactivity did not appear to be mediated by vascular endothelial growth factor (VEGF) because stimulation had no measurable effect on VEGF or VEGF-receptors levels (Supplementary Fig. 8b–d).

Our data show that perturbation of neural activity reduces vascular growth by 70–80% following activity paradigms and near 100% following seizures. This growth arrest is likely to affect micro-regional tissue oxygenation because the brain, unlike other organs, has high energetic demand, minimal energy storage, and no un-perfused microvascular reserve[1]. To determine the effects of vessel reductions on brain homeostasis, we devised a method to spatially correlate regional tissue oxygenation with the relative proximity to capillaries using the tissue hypoxia probe pimonidazole. At normal oxygen levels, no pimonidazole signal was observed in either control or hyperstimulated mice (Supplementary Fig. 9). Interestingly, when mice were exposed to 8% O_2 for 1 h, we observed a greater pimonidazole signal in mice with reduced microvascular density caused by hyperactivity (Fig. 4a–d). Brain microregions distant from vessels experience lower oxygen tension at baseline and following neural activation[23,24]. Consistent with this, in control mice we observed a trend towards increased pimonidazole signal in areas further than 10 μm from blood vessels. In hyperstimulated mice this difference was more pronounced, suggesting that reduced microvascular density impairs oxygen delivery to areas distant from capillaries (Supplementary Fig. 10).

To explore the effect of this diminished oxygen delivery on neuronal connectivity we measured dendritic spine density of layer V pyramidal neurons after exposure to moderate hypoxia (8% O_2 for 48 h). Mice

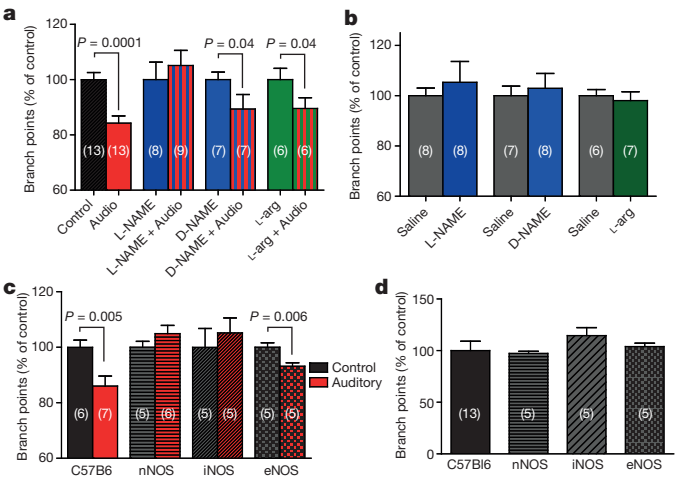


Figure 3 | Inhibition of neuronal and inducible nitric oxide prevents vascular growth arrest in response to activity. **a**, Mice treated with a NOS inhibitor (L-NAME) before daily audio stimulation sessions do not experience vessel reductions following auditory stimulation between P15 and P25; however, pre-treatments with D-NAME and L-arginine do not rescue vessel growth-arrest. **b**, Treatment with L-NAME, D-NAME and L-arginine in the absence of stimulation does not affect cortical vascular density. **c**, Mice deficient in nNOS or iNOS do not experience vessel growth-arrest following auditory stimulation. However, eNOS-deficient mice do experience growth arrest. **d**, Baseline vessel density of all three NOS-deficient strains is not altered compared to wild type. *P* values, one-tailed student's *t*-test. Bars represent s.e.m.

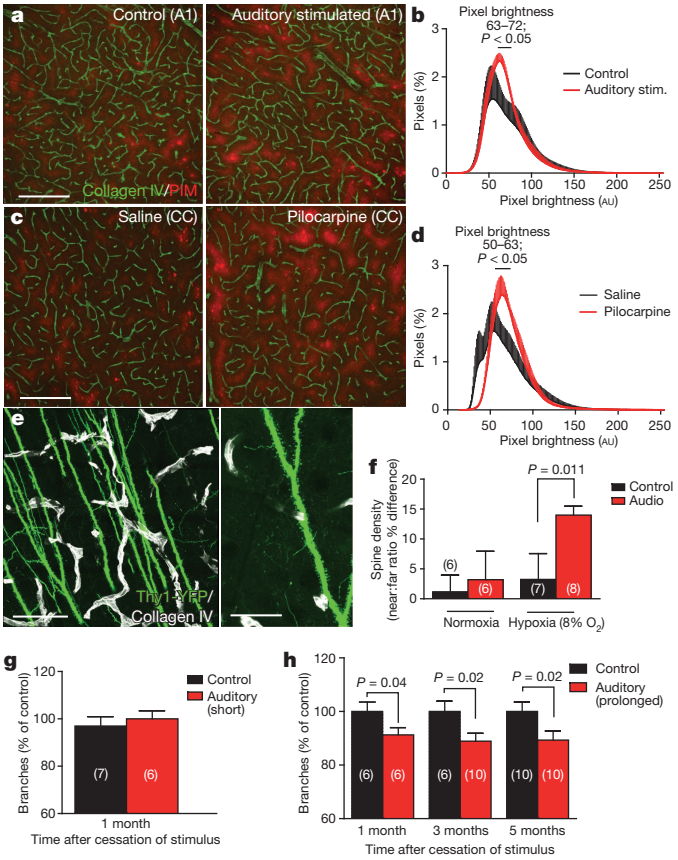


Figure 4 | Activity-mediated microvascular deficits are long-lasting and affect brain oxygenation and dendritic spine stability in areas distant from capillaries. **a**–**d**, Activity-mediated vessel reductions lead to increased levels of hypoxia both in auditory-stimulated (**a**) and pilocarpine-injected (**c**) neonatal mice. Collagen IV vessel label (green) and pimonidazole tissue hypoxia label (red) in primary auditory cortex (A1) of control and auditory-stimulated mice (**a**). Averaged distribution of pimonidazole pixel intensities shows brighter pixels in the stimulated group (**b**) (the average percentage pixels at each pixel intensity were compared between animals for each treatment group using a *t*-test; $n = 5$ control, $n = 4$ auditory stimulated). **c**, Representative images from the cingulate cortex (CC) of saline and pilocarpine-injected mice. **d**, Distribution of pimonidazole brightness shows increased labelling in mice stimulated with pilocarpine ($n = 5$ animals per treatment group). AU, arbitrary units. **e**, **f**, Dendritic spine density in layer 2/3 and 4 of auditory cortex in Thy1-YFPh mice, expressing fluorescent protein in layer 5 pyramidal neurons, was quantified using high-resolution confocal images. Spines per unit length (μm) of dendrite were calculated in areas closer and further than 20 μm from the nearest vessel. **e**, Left panel, projecting dendrites (green) and vessels (white). Right panel, a dendrite studded with spines. **f**, Control and audio-treated mice were left in normoxia or exposed to 8% oxygen for 48 h and spine density was quantified. **g**, **h**, Vascular deficits following a 5-day auditory stimulation are fully recovered 1 month after the stimulus ended (**g**); however, auditory stimulation for at least 15 days leads to permanent vessel loss up to 5 months post stimulation (**h**). Scale bars, 200 μm (**a**, **c**), 100 μm, inset 50 μm (**e**). Bars represent s.e.m.

that had previously been exposed to repetitive auditory stimuli experienced a reduction in spine density in auditory cortex, specifically in areas distant from capillaries, whereas unstimulated mice did not experience spine loss (Fig. 4e, f). This demonstrates that the observed microvascular reductions can cause synaptic loss after a mild reduction in environmental oxygen and suggests that other metabolic challenges or increases in demand may have similar effects. In addition to changes in spine number, our results suggest that impaired oxygen delivery to intercapillary regions may be involved in the alterations in cortical tonotopic mapping and auditory discrimination observed after repetitive auditory stimulation in neonates[25–27].

We next sought to determine whether microvascular reductions were long-lasting by exposing mice to auditory stimulation and assessing vessel density at various time intervals. Neonates given a 5-day auditory stimulus completely regained normal vascular density 1 month later (Fig. 4g), probably because hyperstimulation was stopped before the end of the first postnatal month, when substantial angiogenesis is still ongoing[1]. However, when auditory stimulation was extended to at least 15 days, vessel density did not return to normal even after 5 months (Fig. 4h). This indicates the existence of a critical period during which the vasculature can recover from anti-angiogenic intervention, but after which deficits become permanent. Therefore a temporary exposure to hyperactivity in neonates causes permanent alterations in microvascular architecture, leading to deficits in oxygen delivery and impaired neuronal connectivity (Supplementary Fig. 11).

Our study shows that modest alterations in baseline levels of neural activity do not affect vascular patterning, suggesting that in the postnatal period brain angiogenesis follows an autonomous developmental program. Unexpectedly, this program can be disrupted by increased levels of nitric oxide released from neurons and glia following repetitive sensory-motor stimulation or seizures. This effect is probably maladaptive, given that such repetitive activity patterns are not likely to have been prevalent through evolution. However, these findings raise the concern that early childhood seizures[28] or exposure to repetitive auditory and other sensory-motor stimuli, which are common in modern society[29], could have lifelong repercussions on the cortical microvasculature, its oxygen delivery capabilities, and the homeostasis of neural cells. In addition, it may make the brain vulnerable to conditions of reduced oxygen supply or microvascular pathology such as hypertension, diabetes and ageing.

METHODS SUMMARY

Neonatal or adult mice were exposed daily to various sensorimotor stimulation paradigms (auditory stimulation, whisker deflection, treadmill exercise or chemically induced seizures) for various time intervals. High-resolution confocal microscopy of vascular and proliferation markers or *in vivo* two photon microscopy in mice expressing GFP in endothelial cells was performed to quantify vascular density, proliferation, and rates of vessel formation and elimination. Reduction of nitric oxide synthase (NOS) was achieved by injection of L-NAME or in mice

lacking nNOS, iNOS or eNOS. Measurement of the effects of microvascular reductions on tissue oxygenation was accomplished by microregional densitometric analysis of pimonidazole fluorescence using custom-made ImageJ-based macros. Dendritic spines were quantified in regions near or far from microvessels using confocal microscopy of mice expressing yellow fluorescent protein (YFP) under a neuronal promoter. A full description of methods is available in the online version of the paper.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 January; accepted 29 October 2013.
Published online 4 December 2013.

1. Harb, R., Whiteus, C., Freitas, C. & Grutzendler, J. *In vivo* imaging of cerebral microvascular plasticity from birth to death. *J. Cereb. Blood Flow Metab.* **33,** 146–156 (2013).
2. Spitzer, N. C. Electrical activity in early neuronal development. *Nature* **444,** 707–712 (2006).
3. Lam, C. K., Yoo, T., Hiner, B., Liu, Z. & Grutzendler, J. Embolus extravasation is an alternative mechanism for cerebral microvascular recanalization. *Nature* **465,** 478–482 (2010).
4. Carmeliet, P. & Tessier-Lavigne, M. Common mechanisms of nerve and blood vessel wiring. *Nature* **436,** 193–200 (2005).
5. Dunning, H. S. & Wolff, H. G. The relative vascularity of various parts of the central and peripheral nervous system of the cat and its relation to function. *J. Comp. Neurol.* **67,** 433–450 (1937).
6. Black, J. E., Isaacs, K. R., Anderson, B. J., Alcantara, A. A. & Greenough, W. T. Learning causes synaptogenesis, whereas motor activity causes angiogenesis, in cerebellar cortex of adult rats. *Proc. Natl Acad. Sci. USA* **87,** 5568–5572 (1990).
7. Wei, L., Erinjeri, J. P., Rovainen, C. M. & Woolsey, T. A. Collateral growth and angiogenesis around cortical stroke. *Stroke* **32,** 2179–2184 (2001).
8. Whitaker, V. R., Cui, L., Miller, S., Yu, S. P. & Wei, L. Whisker stimulation enhances angiogenesis in the barrel cortex following focal ischemia in mice. *J. Cereb. Blood Flow Metab.* **27,** 57–68 (2007).
9. Rao, S. *et al.* A direct and melanopsin-dependent fetal light response regulates mouse eye development. *Nature* **494,** 243–246 (2013).
10. Vasudevan, A., Long, J. E., Crandall, J. E., Rubenstein, J. L. & Bhide, P. G. Compartment-specific transcription factors orchestrate angiogenesis gradients in the embryonic brain. *Nature Neurosci.* **11,** 429–439 (2008).
11. Margolis, D. J. *et al.* Reorganization of cortical population activity imaged throughout long-term sensory deprivation. *Nature Neurosci.* **15,** 1539–1546 (2012).
12. Zuo, Y., Yang, G., Kwon, E. & Gan, W.-B. Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* **436,** 261–265 (2005).
13. Crochet, S. & Petersen, C. C. H. Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nature Neurosci.* **9,** 608–610 (2006).
14. Beloozerova, I. N., Sirota, M. G. & Swadlow, H. A. Activity of different classes of neurons of the motor cortex during locomotion. *J. Neurosci.* **23,** 1087–1097 (2003).
15. Rudic, R. D. *et al.* Direct evidence for the importance of endothelium-derived nitric oxide in vascular remodeling. *J. Clin. Invest.* **101,** 731–736 (1998).
16. Keilhoff, G. *et al.* Patterns of nitric oxide synthase at the messenger RNA and protein levels during early rat brain development. *Neuroscience* **75,** 1193–1201 (1996).
17. Buskila, Y. & Amitai, Y. Astrocytic iNOS-dependent enhancement of synaptic release in mouse neocortex. *J. Neurophysiol.* **103,** 1322–1328 (2010).
18. van den Tweel, E. R. W. *et al.* Expression of nitric oxide synthase isoforms and nitrotyrosine formation after hypoxia-ischemia in the neonatal rat brain. *J. Neuroimmunol.* **167,** 64–71 (2005).
19. Ridnour, L. A. *et al.* Nitric oxide regulates angiogenesis through a functional switch involving thrombospondin-1. *Proc. Natl Acad. Sci. USA* **102,** 13147–13152 (2005).
20. Heller, R., Polack, T., Gräbner, R. & Till, U. Nitric oxide inhibits proliferation of human endothelial cells via a mechanism independent of cGMP. *Atherosclerosis* **144,** 49–57 (1999).
21. Jones, M. K., Tsugawa, K., Tarnawski, A. S. & Baatar, D. Dual actions of nitric oxide on angiogenesis: possible roles of PKC, ERK, and AP-1. *Biochem. Biophys. Res. Commun.* **318,** 520–528 (2004).
22. Perrenoud, Q. *et al.* Characterization of type I and type II nNOS-expressing interneurons in the barrel cortex of mouse. *Front. Neural Circuits* **6,** 36 (2012).
23. Kasischke, K. A. *et al.* Two-photon NADH imaging exposes boundaries of oxygen diffusion in cortical vascular supply regions. *J. Cereb. Blood Flow Metab.* **31,** 68–81 (2011).
24. Devor, A. *et al.* ''Overshoot'' of O2 is required to maintain baseline tissue oxygenation at locations distal to blood vessels. *J. Neurosci.* **31,** 13676–13681 (2011).
25. Chang, E. F. & Merzenich, M. M. Environmental noise retards auditory cortical development. *Science* **300,** 498–502 (2003).
26. Zhang, L. I., Bao, S. & Merzenich, M. M. Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nature Neurosci.* **4,** 1123–1130 (2001).
27. Strata, F. *et al.* Perinatal anoxia degrades auditory system function in rats. *Proc. Natl Acad. Sci. USA* **102,** 19156–19161 (2005).
28. Sillanpää, M., Jalava, M., Kaleva, O. & Shinnar, S. Long-term prognosis of seizures with onset in childhood. *N. Engl. J. Med.* **338,** 1715–1722 (1998).
29. Committee. on Environmental Health Noise: a hazard for the fetus and newborn. *Pediatrics* **100,** 724–727 (1997).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors appreciate the expert advice of W. Sessa, M. Simons and F. Moraes. A. Schain helped with design of ImageJ macros and G. P. Flowers critically read the manuscript. This study was supported by the following Grants: R01AG027855 and R01HL106815 (J.G.); F31NS068041 (C.W.) and AHA# 10POST2570007 (C.F.).

Author Contributions C.W. and J.G. conceived the project, C.W., C.F. and J.G. designed the experiment, C.W. and C.F. carried out the experiment, C.W., C.F. and J.G. analysed the data, and C.W. and J.G. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.G. (jaime.grutzendler@yale.edu).

METHODS

Mice. Wild-type mice (Charles River), Tie2-GFP endothelial reporter (Jackson Laboratory no. 003658),Thy1-YFP (Jackson Laboratory no. 003782), neuronal nitric oxide synthase knockdown mice (Jackson Laboratory no. 002986), inducible NOS knockout mice (Jackson Laboratory no. 002609), and endothelial NOS knockout mice (Jackson Laboratory no. 002684) were on a C57Bl6 background. Mice between postnatal age P5 and P200 were used. In all experiments littermates were assigned to treatment or control groups and equally distributed by sex and bodyweight (although neither of these factors influence vessel density; Supplementary Fig. 12d, e). Pups were housed jointly with their mother except for the duration of stimulation or injections. For adult experiments 9-week-old females were used. Sample sizes for each experiment were chosen based on the size of the litter (typically $n = 5$ per group) and in some cases repeated to confirm with different litters. Experimental protocols were in accordance with the relevant guidelines and regulations of the Institutional Animal Care and Use Committee at Northwestern University and Yale University.

Treadmill exercise, auditory and whisker stimulations. In our exercise experiments, pups starting at age P15 were placed in individual lanes on a custom-made treadmill to induce moderate running speeds (15 cm s^{-1}) over 45 min, 3 times daily for 5 days.

For auditory experiments, pups were separated from their mothers and exposed overnight to 10 h of a variety of sounds (40–75 dB) consisting of white and pink noise, tones over a range of frequencies, frequency sweeps, recordings of rodent vocalizations, and other natural sounds. Littermate controls were also separated from their mothers and kept in a quiet room. Stimulation began as early as P15 and lasted 5, 10 or 15 days. For long-term experiments, mice were stimulated between 15 and 30 days and then returned to their home cages for 1, 3 or 5 months before being euthanized.

To stimulate whiskers unilaterally, unilateral whisker trimming was performed daily and mice were subsequently placed into a cage with a continuous airflow for 10 h over 8 days. Vessel density in the stimulated barrel cortex (contralateral to the stimulated whiskers) was compared to the unstimulated (trimmed) cortical region.

Pilocarpine-induced seizures. Pilocarpine was administered intraperitoneally (72 mg kg^{-1}, P6503, Sigma) as early as P5. This dosage is substantially lower than that used to induce chronic epilepsy[30]. Starting at P7 mild seizures were observed following injection. These were characterized by reduced exploratory behaviour, excessive salivation, forelimb clonus and occasional head bobbing, which ceased after 2 h. Mice received daily injections over 10 days. No evidence of spontaneous seizures was noted during this period.

Intracortical tetanus and botulinum toxin injections. For intracortical injections, a craniotomy is made over the cingulate cortex of anaesthetized mice by thinning the skull with a dental drill and using a fine needle to remove a small (100 μm diameter) piece of skull. Using a fine glass capillary mounted to a stereotaxic apparatus, tetanus toxin[31] (25 ng; 190A, List Biological Labs) or botulinum toxin (0.03 ng; 128A, List Biological Labs) was diluted in 7 μl of artificial cerebral spinal fluid (ACSF) and was injected 500 μm deep into the cingulate cortex. A CamKII-Tomato adeno-associated reporter virus was added to the injection mix to demarcate the injected area for later relocation. An equivalent volume of ACSF with virus was injected into control mice. Mice experienced spontaneous seizures, demonstrated by head bobbing and facial spasms a few days after injection, which continued to occur until they were sacrificed 10 days later.

Models of impaired metabolism and stress and nitric oxide manipulations. Starting at P15 mice were given 10 daily intraperitoneal injections of either 2-deoxyglucose (2-DG, 10 mg kg^{-1}; D8375, Sigma), which impairs glucose metabolism, or the glucocorticoid dexamethasone (5 mg kg^{-1}, D4902, Sigma), which is a widely used model of stress[32]. To manipulate nitric oxide levels, the nitric oxide inhibitor L-NAME (50 mg kg^{-1}, N5751, Sigma), the nitric oxide substrate L-arginine[33] (50 mg kg^{-1}, A5006, Sigma), and the L-NAME enantiomer D-NAME (50 mg kg^{-1}, N4770, Sigma) were diluted in saline and injected intraperitoneally 30 min before onset of each stimulation session.

***In vivo* transcranial two-photon time-lapse imaging.** Cerebral blood vessels were imaged in Tie2-GFP endothelial reporter mice. Neonatal time lapse imaging was performed on the two-photon microscope as previously described[1]. In brief, mice were anaesthetized with isoflurane and the skull was exposed with a midline scalp incision. A 1-mm-diameter skull region over the somatosensory cortex was thinned with a microsurgical blade to a final thickness of approximately 30 μm. The surrounding skull was attached to a custom-made steel plate to stabilize the head while imaging. A CCD camera image of skull blood vessels allowed for re-location of the imaging area. The first time point was acquired at P15, and then mice received daily intraperitoneal injections of pilocarpine or saline for 5 days. One day after the final injection, mice were prepared for imaging and blood vessels were re-located and imaged for a second time point.

Hypoxia stress test. Mice were exposed to 15 days of auditory stimulus or 10 days of pilocarpine seizures. Mice were then injected intraperitoneally with Hypoxyprobe-1 (pimonidazole Hcl 50 mg kg^{-1}; HP1-1000Kit, Hypoxyprobe) to label hypoxic tissue. Note that the threshold for pimonidazole detection is <10 mm Hg tissue O_2 and may therefore not detect mild O_2 reductions. Ten minutes after injection of the probe, mice were placed into a hypoxic chamber set at 8% oxygen for 1 h. Mice were anaesthetized, while still in the hypoxic chamber and, when they were no longer responsive, they were removed and immediately euthanized by transcardial perfusion.

Imaging of dendritic spines in hypoxic brains. Dendritic spines were visualized in Thy1-YFPh-line reporter mice, which express yellow fluorescent protein in layer V pyramidal projection neurons. Mice were given 10 days of auditory stimulation and then placed in a hypoxic chamber at 8% for 48 h. Control littermates received no stimulation but were exposed to hypoxia. Animals were injected intravenously with NHS biotin to label vessels and sacrificed by transcardial perfusion. In a separate group of mice, animals were exposed to auditory stimulation or control conditions and immediately sacrificed to determine baseline changes in spine density.

Tissue collection and histology. Mice were anaesthetized by intraperitoneal injection with ketamine–xylazine (120 mg ml^{-1}:10 mg ml^{-1}) and euthanized by transcardial perfusion using 4% paraformaldehyde (PFA). Brains were post-fixed, bathed in sucrose and cut on the cryostat into 50-μm coronal sections. For western blot analysis, fresh cortices were dissected from the rest of the brain under a dissecting microscope and flash frozen. For brain weight measurements, un-perfused brains were extracted and immediately weighed.

Vessels were stained in fixed tissue using collagen IV antibody (1:250; -ab19808, Abcam), Glut-1 (1:100; 2186307, Millipore) or isolectin B4 (1:50; B1205, Vector Labs) as previously described[1]. Additional vessel labelling was achieved by transcardial injection of fluorescein-tomato-lectin (200 μl; FL1171, Vector Labs) or NHS-biotin (300 mg kg^{-1}; 2027, Thermo), which required a conjugated streptavidin counterstain.

For BrdU labelling, mice were injected intraperitoneally every second day with BrdU (5 mg kg^{-1}; B5002, Sigma). Brain sections were washed, denatured with 5M HCl for 15 min and incubated with rat anti-BrdU (1:400, OBT0030, AbDSerotec).

Staining for other cell types was performed using the following antibodies: NeuN (1:250; MAB377, Millipore) to label neurons, IBA1 (1:200; 019-19741, Wako) for microglia, Pdgfrβ (1:100; AF1042, R&D) for pericytes, VEGF (1:100; ab46154, Abcam). To label VEGF pathway components VEGFR2 (1:1,000; 55B11 Cell Signaling), Phospho-VEGFR2 (Tyr 1175) (1:1,000; 19A10, Cell Signaling), Phospho-VEGFR2 (Tyr951) (1:1000; 7H11, Cell Signaling), and VEGFR1 (1:1000; ab32152, Abcam) were used. Hypoxyprobe-1 (1:50, HP1-1000Kit, Hypoxyprobe) was used to detect injected pimonidazole, and c-Fos (1:100; F7799, Sigma) was used to detect immediate early protein detection. DAPI (1:1,000; 9542, Sigma) was used to quantify cellular densities.

A TUNEL detection kit (11684817910, Roche) and cleaved caspase-3 antibody (1:100, Millipore; 559565, BD Pharmingen) were used to label dying and apoptotic cells. DNase treatment served as a positive control for TUNEL staining and a stroke induced by microembolism with 50-μm beads injected into the carotid artery was a positive control for both TUNEL and caspase-3 labelling. AlexaFluor antibodies 488 and 555 (Life Sciences) were used throughout. For all experiments, sections from both groups were stained simultaneously, using the same batch of antibodies.

VEGF protein levels were measured with mouse VEGF Quantikine Elisa Kit (MMV00, R&D) according to the manufacturer's instructions using tissue from mice treated with pilocarpine or saline between P15 and P25 and assayed in triplicate.

Area selection and image acquisition for vascular quantification. Vascular density varies greatly by brain region, therefore the Paxinos and Allen brain atlas were used to carefully identify and locate cortical areas. For each area analysed, a standard coronal section, containing the relevant cortical region, was found. Three consecutive sections were selected from each mouse brain region and standardized regions were located bilaterally using the atlas (six regions of interest were obtained for quantification in each brain, see Supplementary Fig. 12a). To ensure uniformity of the sections, special care was taken that the angle of cryo-cutting was standard for all brains and cutting was performed during a single session.

Primary auditory cortex, sensory barrel cortex and primary motor cortex were quantified for auditory, whisker and motor experiments, respectively. Control areas were always chosen from the same section: cingulate cortex used as a control for auditory and whisker experiments and the piriform cortex used for motor stimulation. For pilocarpine seizures, 2DG-, and dexamethasone-treated brains, the cingulate cortex was quantified.

Using a Leica 0.8 NA 20× lens, an image of the entire cortical region of interest was obtained on a confocal microscope (Leica SP5) at ×1 zoom, 1 μm step size,

1,024 × 1,024 resolution, and a scan speed of 600 Hz. Imaging for an individual experiment was performed in one session, with uniform laser intensity and gain parameters used between control and experimental groups.

Vessel density quantification and laminar layer determination. For quantification, vessels were labelled with isolectinB4, NHS-biotin or collagen IV. Z-projections were made from a fixed number of optical sections. All quantifications were carried out blind. Vascular length was quantified automatically using a custom-made ImageJ (NIH) macro in which vessels were thresholded to create a binary mask, despeckled, and skeletonized. Skeletonized vessels were used to obtain vessel length in microns (Supplementary Fig. 12a). Blinded manual counts of blood vessel branches were performed using Image J Cell Counting software (Supplementary Fig. 12b). To ensure accurate counts, six cortical regions and more than 2,000 branches were counted in each animal. To quantify individual laminar layers, we used previously published measurements of cortical layers in barrel, auditory, and motor cortices[34–36] to divide our images into the relevant laminae.

Other cellular quantifications. We have previously shown that endothelial BrdU can be accurately separated from non-endothelial BrdU as well as peri-vascular BrdU using confocal images[1]. For BrdU, high-resolution confocal images were obtained using a Leica 1.3 40× HCLX PL APO oil immersion lens in the cingulate cortex for pilocarpine-treated mice, and in the auditory and cingulate cortex for auditory-stimulated mice. Z-projections of a standardized thickness were made and BrdU positive endothelial and non-endothelial cells were counted on ImageJ Cell Counting software.

Quantification of total cell number (DAPI), neuron, and microglia number was performed in the cingulate cortex using ImageJ Cell Counting software. All manual counting was carried out blind.

For co-localization of neurons with c-Fos, neurons and c-Fos images were separated and quantified individually on ImageJ Cell counter so that the percentage of c-fos positive neurons could be calculated.

Quantification of microvasculature *in vivo*. *In vivo* two-photon image stacks were cut to a standard thickness of 150 μm and vessels were quantified blind. To measure total vessel length, we used the program Fiji Simple Neurite Tracer, which allowed us to trace individual vessels in three dimensions. Total vessel length was determined for first and final time points and per-cent difference was calculated.

ImageJ was used to quantify elimination and formation of branches *in vivo* by comparing first and second time points. New branches consisted of new sprouts (counted as one new branch), sprouts which had anastomosed to form a connected vessel (counted as one new branch), and vessels which became fully connected at the second time point (counted as two new branches). Eliminated branches consisted of fully formed vessels that had retracted such that only a sprout remained (counted as one eliminated branch), fully formed vessels which retracted completely (counted as two eliminated branches), and sprouts which retracted completely (counted as one eliminated branch). Change in branches was determined by counting all branch points at time point 1 and then counting branches that had been added or lost at time point 2. Per-cent change in branch formation and elimination were calculated by dividing the number of new or lost branches by the total number of branches at time point 1. The observed vessel changes were evenly distributed across the entire tissue volume

Statistics. For each mouse, vessel and cell-density quantifications were obtained from both hemispheres of three coronal sections (six images in total) and were averaged to obtain a single value.

Although there is strong inter-litter consistency in vessel density, there are small but statistically significant differences in density between litters. Therefore all of our comparisons are carried out within litters (both controls and experimental animals belong to a single litter). However, as most experiments use multiple litters, we use a normalization method to add together different litters. In each litter, the experimental group is normalized so that the control group averages to 100%, and following this normalization, multiple litters can be combined. Density values are distributed normally and a one tailed student's *t*-test assuming equal variance was used to compare groups. Spine density and hypoxia were compared between areas near and far from vessels, the stimulated and unstimulated hemispheres were compared in whisker-stimulated mice, and all other comparisons were performed between control and experimental groups.

Tissue-hypoxia quantification. Sections were co-stained with collagen IV and pimonidazole antibodies, and images were obtained on Leica HyD detectors, confocal photon-counting mode using a 1.0 NA 20× lens at ×1 zoom, 1 μm step size, 1,024 × 1,024 resolution, and a scan speed of 600 Hz. Primary auditory cortex was imaged following auditory treatment and cingulate cortex was used for pilocarpine seizure treatments and images were projected to uniform thickness (5 μm).

We designed an ImageJ macro to quantify pimonidazole staining. In this, blood vessels are thresholded and a mask is created and projected over the pimonidazole staining so that only pimonidazole signal in non-vascular areas is quantified. A histogram of the pimonidazole staining is obtained which shows the distribution of the brightness of individual pixels. We use 8-bit images, therefore brightness is distributed over 256 bins, 0 representing no detectable signal and 256 representing the brightest. Because the non-vascular area may vary, brightness values were converted to percentages of the total number of pixels counted. Values for each brightness bin (0–256) are averaged across six images for each mouse, to obtain a single average histogram per animal. Group averages for each brightness bin are obtained by averaging the brightness values of individual animals and one-tailed *t*-tests are performed in each brightness bin to compare mice in different experimental groups.

To determine the differences in pimonidazole brightness near and far from capillaries, a histogram representing tissue closer than 10 μm to vessels was compared to areas more than 10 μm away from the nearest vessel. This was done automatically using a custom-made ImageJ macro. In this macro a vessel mask was created and a histogram of total pimonidazole staining was obtained. Then the vessel mask was dilated to 10 μm to obtain a second histogram in the area further from vessels. This second histogram was subtracted from the initial histogram to obtain a value for the area closer than 10 μm to vessels.

Dendritic-spine-density quantification. NHS biotin-injected brains were counterstained with fluorescently labelled streptdavidin to label vessels and sections were imaged on the confocal microscope using a Leica 1.4 NA 63× lens at ×4 zoom at 1 μm step size. Sixteen images (1,024 × 1,024 pixel each) were acquired to cover layers 2/3 and 4 of the auditory cortex and stitched together to reconstruct the dendritic projections. Tiled stacks were Z-projected to a standardized thickness (5 μm) for quantification and four to six images were quantified for each animal. Using a custom-made ImageJ macro, vessels were thresholded and a mask was created. The vessel mask was dilated to 20 μm, to create a selection which outlined the area 20 μm from the nearest vessel. This selection was projected over the image of the dendrites to create a border between areas 'near' and 'far'.

Images were quantified blind to treatment. In each image we excluded dendrites that were out of focus or displayed weak YFP labelling, such that individual spines could not be identified. The remaining dendrites were traced to determine their total length in areas 'near' or 'far' from vessels. Spines were counted manually using ImageJ Cell Counter and density 'near' and 'far' from vessels was determined by dividing by the length of the dendrites quantified. An average density 'near' and 'far' was determined for each animal and group averages 'near' and 'far' were compared using a student's *t*-test.

30. Arida, R. M., Scorza, F. A., de Araujo Peres, C. & Cavalheiro, E. A. The course of untreated seizures in the pilocarpine model of epilepsy. *Epilepsy Res.* **34,** 99–107 (1999).
31. Wykes, R. C. *et al.* Optogenetic and potassium channel gene therapy in a rodent model of focal neocortical epilepsy. *Sci. Transl. Med.* **4,** 151ra152 (2012).
32. Sorrells, S. F., Caso, J. R., Munhoz, C. D. & Sapolsky, R. M. The stressed CNS: when glucocorticoids aggravate inflammation. *Neuron* **64,** 33–39 (2009).
33. Palmer, R. M., Ashton, D. S. & Moncada, S. Vascular endothelial cells synthesize nitric oxide from L-arginine. *Nature* **333,** 664–666 (1988).
34. Lefort, S., Tomm, C., Floyd Sarria, J.-C. & Petersen, C. C. H. The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron* **61,** 301–316 (2009).
35. Oviedo, H. V., Bureau, I., Svoboda, K. & Zador, A. M. The functional asymmetry of auditory cortex is reflected in the organization of local cortical circuits. *Nature Neurosci.* **13,** 1413–1420 (2010).
36. Saikali, S. *et al.* A three-dimensional digital segmented and deformable brain atlas of the domestic pig. *J. Neurosci. Methods* **192,** 102–109 (2010).

LETTER

doi:10.1038/nature12807

Muc5b is required for airway defence

Michelle G. Roy[1]*, Alessandra Livraghi-Butrico[2]*, Ashley A. Fletcher[3]*, Melissa M. McElwee[1], Scott E. Evans[1], Ryan M. Boerner[4], Samantha N. Alexander[1], Lindsey K. Bellinghausen[1], Alfred S. Song[1], Youlia M. Petrova[1], Michael J. Tuvim[1], Roberto Adachi[1], Irlanda Romo[1,5], Andrea S. Bordt[6], M. Gabriela Bowden[6,7], Joseph H. Sisson[8], Prescott G. Woodruff[9], David J. Thornton[10], Karine Rousseau[10], Maria M. De la Garza[1], Seyed J. Moghaddam[1], Harry Karmouty-Quintana[4], Michael R. Blackburn[4], Scott M. Drouin[4], C. William Davis[2], Kristy A. Terrell[2], Barbara R. Grubb[2], Wanda K. O'Neal[2], Sonia C. Flores[3], Adela Cota-Gomez[3], Catherine A. Lozupone[3], Jody M. Donnelly[3], Alan M. Watson[3], Corinne E. Hennessy[3], Rebecca C. Keith[3], Ivana V. Yang[3], Lea Barthel[3,11], Peter M. Henson[3,11], William J. Janssen[3,11], David A. Schwartz[3], Richard C. Boucher[2], Burton F. Dickey[1] & Christopher M. Evans[1,3]

Respiratory surfaces are exposed to billions of particulates and pathogens daily. A protective mucus barrier traps and eliminates them through mucociliary clearance (MCC)[1,2]. However, excessive mucus contributes to transient respiratory infections and to the pathogenesis of numerous respiratory diseases[1]. *MUC5AC* and *MUC5B* are evolutionarily conserved genes that encode structurally related mucin glycoproteins, the principal macromolecules in airway mucus[1,3]. Genetic variants are linked to diverse lung diseases[4–6], but specific roles for MUC5AC and MUC5B in MCC, and the lasting effects of their inhibition, are unknown. Here we show that mouse Muc5b (but not Muc5ac) is required for MCC, for controlling infections in the airways and middle ear, and for maintaining immune homeostasis in mouse lungs, whereas Muc5ac is dispensable. Muc5b deficiency caused materials to accumulate in upper and lower airways. This defect led to chronic infection by multiple bacterial species, including *Staphylococcus aureus*, and to inflammation that failed to resolve normally[7]. Apoptotic macrophages accumulated, phagocytosis was impaired, and interleukin-23 (IL-23) production was reduced in *Muc5b*$^{-/-}$ mice. By contrast, in mice that transgenically overexpress *Muc5b*, macrophage functions improved. Existing dogma defines mucous phenotypes in asthma and chronic obstructive pulmonary disease (COPD) as driven by increased MUC5AC, with MUC5B levels either unaffected or increased in expectorated sputum[1,8]. However, in many patients, MUC5B production at airway surfaces decreases by as much as 90%[9–11]. By distinguishing a specific role for Muc5b in MCC, and by determining its impact on bacterial infections and inflammation in mice, our results provide a refined framework for designing targeted therapies to control mucin secretion and restore MCC.

Mucosal surfaces are central interfaces between organisms and their external environments. Mucus-coated barriers defend against pathogens[12] and re-distributed commensal organisms[13]. Gastrointestinal mucins prevent *Helicobacter pylori* growth[12], colitis[14] and colorectal carcinogenesis[15]. To test whether secreted airway mucins serve correspondingly important roles, we examined MCC and responses to bacterial infections in *Muc5ac*$^{-/-}$ mice (ref. 16), *Muc5b*$^{-/-}$ mice, and lung specific-Muc5b-overexpressing transgenic (Tg(*Scgb1a1-Muc5b*)) mice (Extended Data Fig. 1a–d). We identified unique mechanisms by which Muc5b mediates effective respiratory mucosal defence (Extended Data Fig. 2). In *Muc5b*$^{-/-}$ upper airways, olfactory gland glycoconjugates were absent, but nasopharyngeal surfaces were unaffected (Fig. 1a, b). Muc5ac and Muc5b were lacking in respective knockout airways, but tracheobronchial glycoconjugates increased in *Muc5b*$^{-/-}$ mice owing to induced Muc5ac (Fig. 1c, d and Extended Data Figs 1f and 3). Despite retaining mucous phenotypes in many airway tissues, growth and survival were impaired in *Muc5b*$^{-/-}$ animals (Fig. 1e, f), whereas *Muc5ac*$^{-/-}$ and *Scgb1a1-Muc5b* mice survived normally (Fig. 1f). Acute MCC was normal in *Muc5ac*$^{-/-}$ and *Scgb1a1-Muc5b* mice, but severely reduced in *Muc5b*$^{-/-}$ mice (Fig. 1g, h and Extended Data Fig. 1e), even though functional ciliated cells were present (Fig. 1d, i). Mucus transport was impaired in *Muc5b*$^{-/-}$ tracheal epithelial cells *in vitro*, confirming that defective clearance reflected altered mucociliary interactions specifically (Fig. 1j–l and Supplementary Videos 1 and 2). Collectively, these data identify non-redundant protective requirements for Muc5b in survival and MCC.

Impaired MCC in *Muc5b*$^{-/-}$ mice was accompanied by abnormal breathing (Fig. 2a, b) and hypoxaemia (Fig. 2c). We assessed lung function in mechanically ventilated mice. Baseline airflow and responses to the bronchoconstricting agent methacholine were normal in the lower airways (Fig. 2d–f). Circumventing the upper airways restored ventilation in spontaneously breathing trachestomized animals (Fig. 2g and Supplementary Videos 3 and 4). Thus, upper respiratory obstruction impeded airflow in *Muc5b*$^{-/-}$ mice. Micro-computed tomography (micro-CT) confirmed this with radiological evidence of upper airway obstruction (Fig. 2h), and middle-ear effusion consistent with otitis media (Extended Data Fig. 4a, b). The latter was unexpected given associations between increased MUC5B and human otitis media [17]. In *Muc5b*$^{-/-}$ mice, but not *Muc5ac*$^{-/-}$ mice (data not shown), hair fragments encased in mucus-like material were consistently found in posterior nasopharynxes (Fig. 2i, j and Extended Data Fig. 4c) and middle ears (Fig. 3a and Extended Data Fig. 4d). Bacteria and inflammation in *Muc5b*$^{-/-}$ middle-ear lavage samples confirmed infectious otitis media (Fig. 3b, c).

In *Muc5b*$^{-/-}$ lower airways, aspirated materials and inflammatory infiltrates were also common (Fig. 3d, e). Culturable bacteria in the lungs increased 2.9–21.6-fold over time, and 7.6–75-fold further in spontaneously moribund mice who also had increased bacteria in spleen cultures (Fig. 3f, g), suggesting that disseminated infections contributed to mortality. To test this hypothesis, mice were placed on antibiotic-supplemented diets. Antibiotics reversed spontaneous lethality and reduced lung bacterial burden, but did not restore normal ventilation in *Muc5b*$^{-/-}$ mice (Fig. 3h–j). Thus, the cause of death in *Muc5b*$^{-/-}$ mice was infectious, and not directly due to airflow limitation. These data identify a natural course in which *Muc5b* deficiency causes upper airway obstruction, spontaneous infection of connecting auditory tubes and the lungs, and fatal bacteraemia.

[1]University of Texas, MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. [2]University of North Carolina-Chapel Hill, 7011 Thurston-Bowles Building, Chapel Hill, North Carolina 27599, USA. [3]University of Colorado School of Medicine, 12700 East 19th Avenue, Aurora, Colorado 80045, USA. [4]University of Texas Health Science Center-Houston Medical School, 6431 Fannin Street, Houston, Texas 77030, USA. [5]Instituto Tecnológico y de Estudios Superiores de Monterrey, Avenida Eugenio Garza Sada 2501 Sur Colonia Tecnológico, Monterrey, Nuevo León 64849, Mexico. [6]Texas A&M Health Science Center, 2121 W. Holcombe Boulevard, Houston, Texas 77030, USA. [7]University of Houston-Downtown, 1 Main Street, Houston, Texas 77002, USA. [8]University of Nebraska Medical Center, 985910 Nebraska Medical Center, Omaha, Nebraska 68198, USA. [9]University of California San Francisco, 505 Parnassus Avenue, San Francisco, California 27599, USA. [10]University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK. [11]National Jewish Health, Denver, Colorado 80206, USA.
*These authors contributed equally to this work.

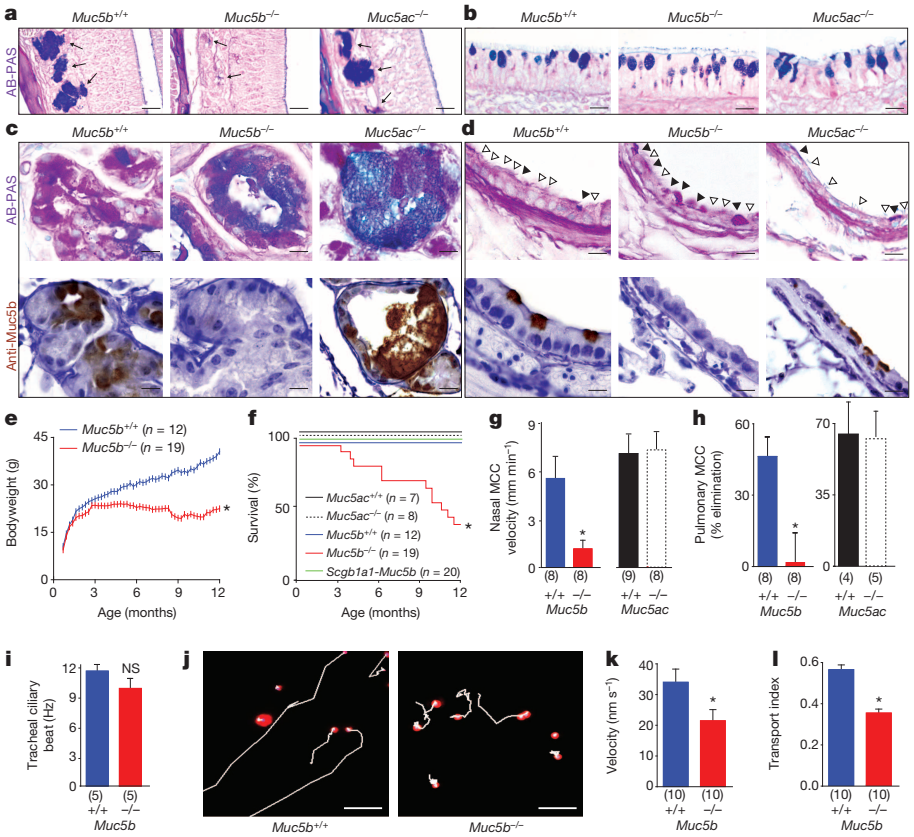
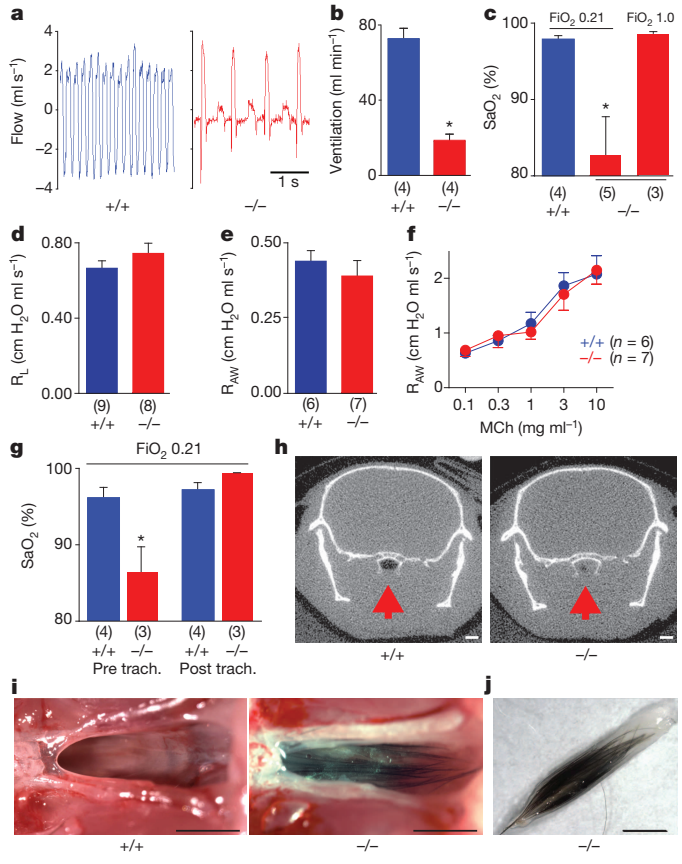


Figure 1 | Muc5b is required for survival and particle clearance. **a–d**, $Muc5b^{+/+}$, $Muc5b^{-/-}$ and $Muc5ac^{-/-}$ olfactory glands (**a**), nasopharynxes (**b**), tracheal glands (**c**), and bronchial surface epithelial cells (**d**) stained with AB-PAS (Alcian blue staining combined with periodic acid–Schiff (PAS) staining) (**c**, **d**, top part) or with anti-Muc5b (**c**, **d**, bottom part). Solid arrowheads, mucous cells. Open arrowheads, ciliated cells. **e–i**, Effects of mucin expression on growth, survival, MCC and ciliary function. **j–l**, Microsphere movement, transport velocity and transport index in primary tracheal epithelial cells *in vitro*. Scale bars, 20 μm (**a**, **j**), 10 μm (**b–d**). Error bars, s.e.m. $*P < 0.05$; NS, not significant. Numbers in brackets, *n* mice.



Because of the importance of bacterial infection in $Muc5b^{-/-}$ mice, we identified organisms whose accumulation in the lungs was normally prevented by Muc5b-rich mucus. 16S rRNA phylotyping revealed 19 genera and >25 species in the lungs. Most were oral, gut and skin microbiota. At 3 months of age, populations were heterogeneous and indistinguishable (Fig. 3k and Extended Data Fig. 5). However, in moribund $Muc5b^{-/-}$ lungs, streptococci and staphylococci increased significantly (Fig. 3k). One streptococcal species, (*Streptococcus acidominimus*) increased in abundance but was also frequently found in controls (Extended Data Fig. 5f–h). By contrast, *Staphylococcus aureus*, an important pneumonia-causing pathogen, selectively increased in spontaneously moribund $Muc5b^{-/-}$ mice (150–389-fold; Fig. 3k). Its incidence was 78% in moribund mice compared to 25% in 3-month-old $Muc5b^{-/-}$ mice, and 8% in $Muc5b^{+/+}$ mice (chi-squared test, $P = 0.002$). Moribund mouse spleen cultures contained bacteria also found in the lungs, including *S. aureus* (Extended Data Fig. 6). Based on these associations, we tested the effects of mucin expression on pathogenesis following airway inoculation with *S. aureus* USA300 (ref. 18). Although this was non-lethal in $Muc5b^{+/+}$, *Scgb1a1-Muc5b*, and $Muc5ac^{-/-}$ animals, only 40% of $Muc5b^{-/-}$ mice survived (Fig. 3l and Extended Data Fig. 7a–e). Thus, Muc5b is selectively required for preventing acquisition and enrichment of virulent bacteria such as *S. aureus* in the airways.

Figure 2 | Muc5b deficiency causes severe upper airway obstruction. **a**, **b**, Airflow and ventilation in $Muc5b^{+/+}$ and $Muc5b^{-/-}$ mice. **c**, Oxygen saturation (SaO_2) and rescue with 100% O_2 supplementation (fraction of inspired oxygen (FiO_2), 1.0). **d–f**, Lung (R_L) and lower airway (R_{AW}) resistance and hyperresponsiveness to methacholine (MCh) in mechanically ventilated mice. **g**, Upper airway bypass by tracheostomy (trach.) under normoxia (FiO_2, 0.21). **h**, Upper airway obstruction in $Muc5b^{-/-}$ mice confirmed by micro-CT (red arrows, nasopharynxes). **i**, **j**, Hair encased in mucus-like plugs visible during necropsy. Scale bars, 1 mm. Error bars, s.e.m. $*P < 0.05$. Numbers in brackets, *n* mice.

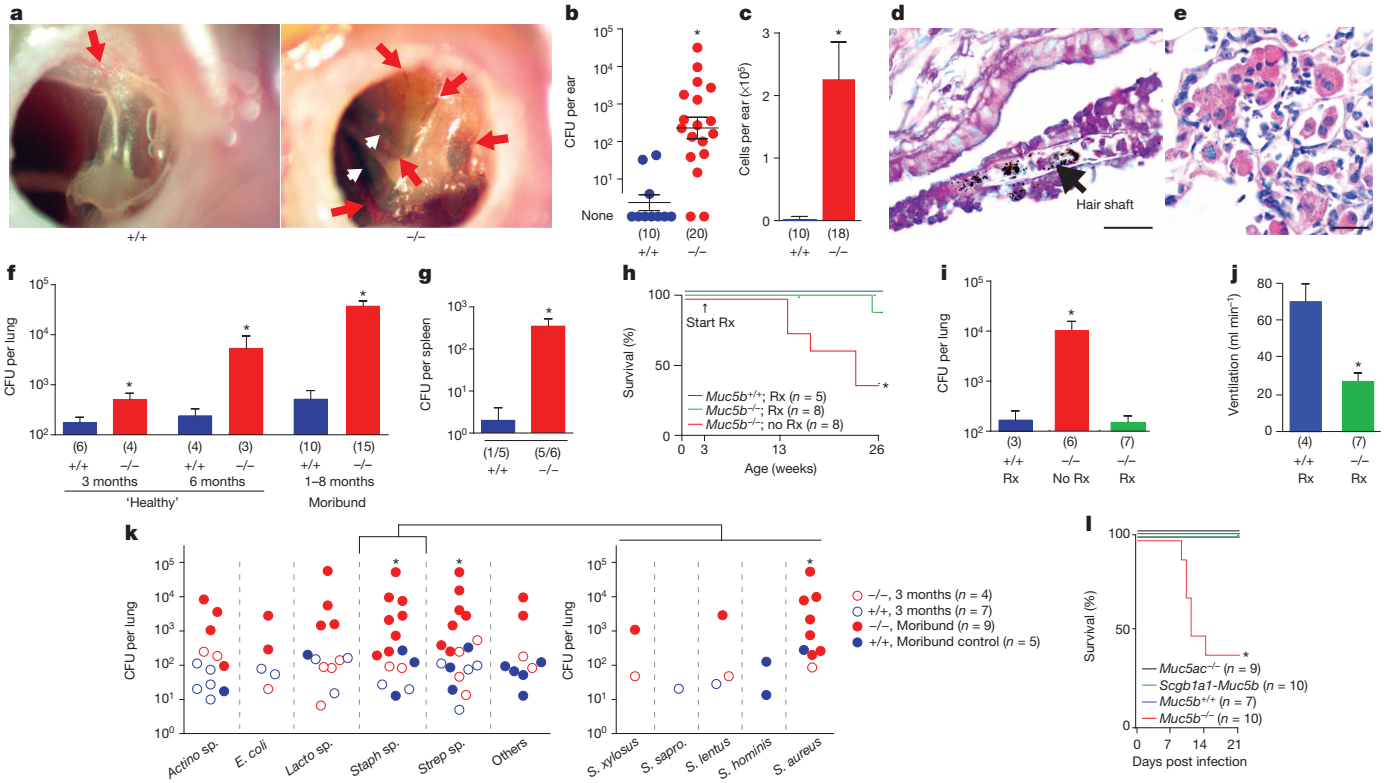


Figure 3 | Infection is the cause of death in *Muc5b*$^{-/-}$ mice. **a**–**c**, Hyperaemic blood vessels (red arrows), hair (white arrows), bacteria and leukocytes (>95% neutrophils) in *Muc5b*$^{-/-}$ middle ears. **d**, **e**, Hair fragments and leukocytes in *Muc5b*$^{-/-}$ lungs. **f**, **g**, Chronically accumulating bacteria in lungs, leading to moribundity and septicemia in *Muc5b*$^{-/-}$ mice. **h**–**j**, Mortality, infection and ventilation with antibiotic treatment. **k**, Bacterial genera, and staphylococcal species in *Muc5b*$^{-/-}$ mice. **l**, Survival following intranasal *S. aureus* inoculation (10^7 CFU per mouse). *Actino*, *Actinobacillus* sp.; *E. coli*, *Escherichia coli*; *Lacto*, *Lactoobacillus* sp.; Rx, antibiotic treatment; *S. aureus, Staphylococcus aureus*; *S. hominis, Staphylococcus hominis*; *S. lentus, Staphylococcus* lentus; *S. sapro, Staphylococcus saprophyticus*; *Staph*, *Staphylococcus*; *Strep*, *Streptococcus*; *S. xylosus*, *Staphylococcus xylosus*. Scale bars, 20 μm (**d**, **e**). Error bars, s.e.m. *$P < 0.05$. Numbers in brackets, *n* mice.

The airways are exposed to bacteria from birth onwards. MCC is considered the major route for their elimination. However, although *Muc5b*$^{-/-}$ mice showed severely impaired MCC, most survived several months with increased lung bacterial burden (Figs 1f and 3f, h), but were ultimately unable to control *S. aureus* infection (Fig. 3k, l). Therefore, we investigated additional potential mechanisms for Muc5b-dependent airway defence. Purified MUC6 (but not MUC5AC) inhibits *H. pylori* growth[12], and purified MUC5AC (but not MUC2) inhibits *Trichuris muris* survival[16]. However, neither MUC5AC nor MUC5B inhibited *S. aureus* growth (Extended Data Fig. 7f), suggesting that Muc5b-mediated defence was not mediated by direct inhibitory interactions with *S. aureus*.

We then investigated whether *Muc5b* deficiency affected inflammatory defences. Leukocytes in *Muc5b*$^{-/-}$ lung lavage fluid were significantly altered. In the steady state, neutrophils and eosinophils accumulated, but lymphocytes were absent (Extended Data Fig. 8a). There were also marked changes in macrophages in *Muc5b*$^{-/-}$ lungs. Macrophages accumulated over time (Fig. 4a). They frequently contained undigested cytoplasmic inclusions (Fig. 4b), and had impaired phagocytic functions (Extended Data Fig. 9a). By age 12 months, IL-23, a mediator of anti-microbial inflammatory responses normally produced by macrophages and dendritic cells, was reduced by 93% in *Muc5b*$^{-/-}$ mice (Fig. 4c and Extended Data Table 1). As *Muc5b*$^{-/-}$ mice became moribund over this timecourse, dysfunctional inflammatory phenotypes progressed (Fig. 4d and Extended Data Fig. 8a). Macrophages that accumulated in airspaces were increasingly apoptotic (Fig. 4e, f). IL-23 remained significantly lower in moribund *Muc5b*$^{-/-}$ mice than controls (Fig. 4g), despite their severely infected states (Fig. 3f, k). We found no evidence that macrophages produce Muc5b (Extended Data Fig. 9b), so it is unlikely that these pathologies were cell autonomous. Rather, the ability of *Muc5b*$^{-/-}$ lungs to resolve inflammation following microbial exposure was significantly inhibited[7].

Apoptotic macrophage accumulation in *Muc5b*$^{-/-}$ mice suggests that in addition to acute microbial elimination, MCC is required for chronically maintaining phagocyte clearance and immune functions *in vivo*. We therefore tested the impacts of age and Muc5b expression on acute and resolving lung inflammation following *S. aureus* infection. In young animals (age 3 months), acute neutrophilic responses were similar irrespective of genotype (Extended Data Fig. 8b). By contrast, ageing *Muc5b*$^{-/-}$ animals (age 6 to 7 months) were exquisitely sensitive, exhibiting 100% mortality within the first few hours of infection, whereas *Muc5b*$^{+/+}$ and *Scgb1a1-Muc5b* mice survived (Fig. 4h). Thus, the combined absence of MCC and the chronic effects of underlying spontaneous infections enhanced the deleterious consequences of *S. aureus* infection. Young mice survived to day 7 post infection, a peak time point of resolving inflammation[7]. Infected *Muc5b*$^{-/-}$ lungs exhibited increases in lung macrophages and mixed granulocytes (Fig. 4h and Extended Data Fig. 8c). Accumulated macrophage pools were predominated by apoptotic cells (Fig. 4i) and linked to significantly reduced lung IL-23 (Fig. 4j), similar to spontaneously moribund *Muc5b*$^{-/-}$ mice. Conversely, in *Scgb1a1-Muc5b* mice, IL-23 production, macrophage activation, and *S. aureus* elimination were enhanced (Fig. 4j–l).

Collectively, these data show that Muc5b expression affects inflammation and macrophage-mediated responses in the lungs. In its absence, mice failed to mount effective anti-bacterial responses. In its sustained presence, these were enhanced. Apoptotic cells, including neutrophils, accumulated in MCC deficient *Muc5b*$^{-/-}$ airspaces (Fig. 4e and Extended Data Fig. 9c). Efferocytosis of neutrophils is reported to downregulate IL-23 (ref. 19), which may explain its decline

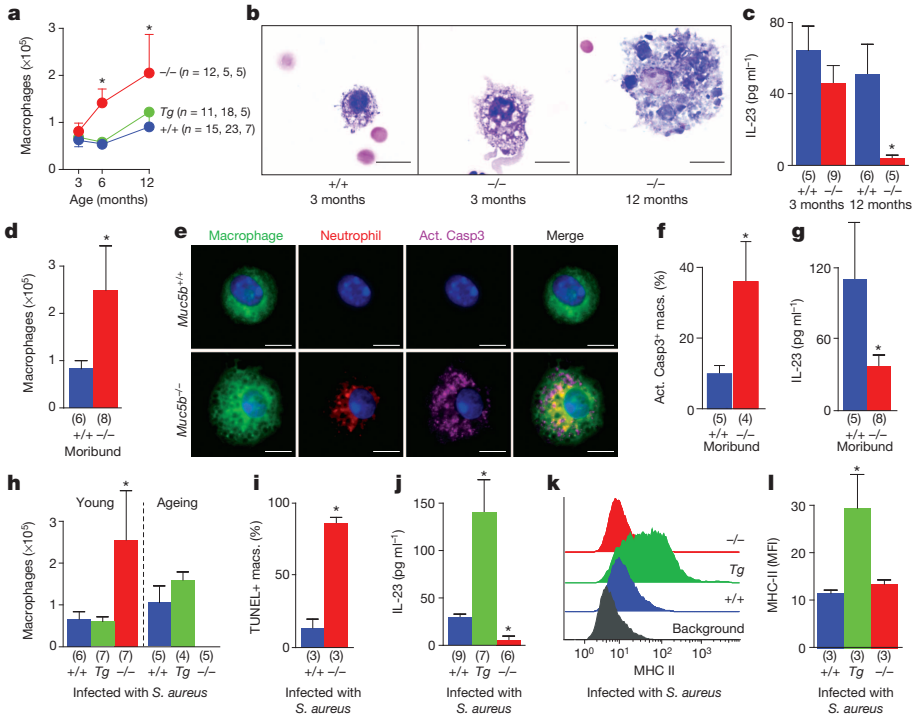


Figure 4 | Muc5b maintains functioning lung macrophage populations. a–c, Temporal changes in lung lavage macrophage numbers and morphologies and in IL-23. **d–g**, Macrophage numbers and apoptosis, and IL-23, in spontaneously moribund and control lungs. **h–j**, Macrophage numbers and apoptosis, and IL-23, in young (3 months) and old (6–7 months) lungs 7 days after intratracheal USA300 infection (10^8 CFU per mouse). Part **h**, empty rightmost bar shows that all mice were dead. **j–l**, IL-23 and macrophage MHC-II in infected *Scgb1a1-Muc5b* mice. *Scgb1a1-Muc5b* lavage USA300 DNA was >98% reduced; incidence was 25% versus 67% ($Muc5b^{+/+}$) and 100% ($Muc5b^{-/-}$), chi-squared test, $P = 0.03$. Error bars, s.e.m. $^*P < 0.05$. Representative data in **k**. Numbers in brackets, *n* mice. Act., activated; macs., macrophages; MHC II, major histocompatability complex class II; *Tg*, *Scgb1a1-Muc5b* mice.

in the lungs. It will be of great interest to test the degrees to which Muc5b-dependent changes in macrophages are mediated by altered elimination owing to impaired MCC, and by other processes such as opsonization or pharmacologic interactions[20]. Nevertheless, these studies functionally link two fundamental first line host defence components; mucus and phagocytes. To our knowledge, Muc5b is the only bronchial secretory protein identified so far that demonstrates such singularly strong biological necessity. Mice survive normally in the absence of Muc5ac or anti-microbial peptides such as β defensin-1 (ref. 21), calgranulin B[22], lipocalin-2 (ref. 23), or lactoferrin[24]. In addition to controlling MCC and macrophage readiness, it is conceivable that Muc5b is a scaffold for these and other defensive molecules.

The novel relationships between Muc5b, MCC and inflammation reported here have the potential to impact therapies for airway diseases. Cystic fibrosis, COPD and asthma are associated with mucus hypersecretion, impaired MCC, and increased risk for respiratory infections[25–29]. *MUC5AC* and *MUC5B* expression varies among individuals, but the overall result is excessive mucin production[1,8–11]. Such pathological associations have overshadowed the benefits of airway mucus and the potential functions of individual mucins in infections[1,12,13,15,16,30]. Human *MUC5B* is highly polymorphic. A recently identified promoter variant found in approximately 20% of the general population increases its expression 37.4-fold in healthy lungs[5]. Our findings in *Muc5b*-deficient and *Muc5b*-overexpressing mice suggest that *MUC5B* variants may regulate airway homeostasis, disease pathogenesis, and mucosal immune function in humans broadly. Although controlling mucin hypersecretion is an attractive therapeutic goal for transient and chronic lung diseases[1], a therapy that completely disrupts mucus and inhibits MUC5B for extended periods may not be advisable. Instead, promoting adequate expression and enhancing MCC and airway defence in controlled manners are better strategies.

METHODS SUMMARY

Mice were used with University of Colorado, University of North Carolina Chapel Hill, and University of Texas M.D. Anderson Cancer Center Institutional Animal Care and Use Committee approvals. $Muc5ac^{-/-}$ mice were generated as described previously[16]. $Muc5b^{-/-}$ and *Scgb1a1-Muc5b* mice were generated as part of the current studies. Muc5b protein was assessed immunohistochemically using rabbit polyclonal antisera. Ciliary beat, MCC, and transport were assessed as described previously. Lung function was measured using a head-out plethysmograph and a flexiVent (Scireq), and blood oxygen was assessed using a pulse oximeter. Otitis media was assessed by visual otoscopy and middle ear lavage (MEL). Pulmonary inflammation was assessed by histology and lung lavage. Lavaged leukocytes were identified by light microscopy and flow cytometry. Neutrophils, macrophages, MHC-II, and apoptotic cells, were detected using commercially available antibodies and reagents. *S. aureus* was administered by 10 µl intranasal or 50 µl intratracheal inocula at 10^7–10^8 colony-forming units per animal. Bacteria and bacterial DNA were isolated from MEL, lung homogenates and lung lavage pellets. Isolated colonies were phylotyped by 16S ribosomal RNA sequencing. Kaplan–Meier (Figs 1f and 3h, l), regression (Figs 1e and 2f), one-sided *t*-test (Figs 1g–i, k, l, 2b–e, g, 3b, c, f, g, j, k and 4c, d, f, g, i, j), and one-way analysis of variance (ANOVA) (Figs 3i and 4a, h, j, l) with appropriate corrections for multiple comparisons, unequal variances, and non-Gaussian distribution were carried out using GraphPad Prism v5.04 (GraphPad Software).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 October 2012; accepted 28 October 2013.
Published online 8 December 2013.

1. Fahy, J. V. & Dickey, B. F. Airway mucus function and dysfunction. *N. Engl. J. Med.* **363,** 2233–2247 (2010).
2. Button, B. *et al.* A periciliary brush promotes the lung health by separating the mucus layer from airway epithelia. *Science* **337,** 937–941 (2012).
3. Young, H. W. *et al.* Central role of *Muc5ac* expression in mucous metaplasia and its regulation by conserved 5′ elements. *Am. J. Respir. Cell Mol. Biol.* **37,** 273–290 (2007).
4. The Collaborative Study on the Genetics of Asthma. A genome-wide search for asthma susceptibility loci in ethnically diverse populations. *Nature Genet.* **15,** 389–392 (1997).
5. Seibold, M. A. *et al.* A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N. Engl. J. Med.* **364,** 1503–1512 (2011).
6. Kamio, K. *et al.* Promoter analysis and aberrant expression of the *MUC5B* gene in diffuse panbronchiolitis. *Am. J. Respir. Crit. Care Med.* **171,** 949–957 (2005).
7. Janssen, W. J. *et al.* Fas determines differential fates of resident and recruited macrophages during resolution of acute lung injury. *Am. J. Respir. Crit. Care Med.* **184,** 547–560 (2011).
8. Thornton, D. J., Rousseau, K. & McGuckin, M. A. Structure and function of the polymeric mucins in airways mucus. *Annu. Rev. Physiol.* **70,** 459–486 (2008).
9. Woodruff, P. G. *et al.* T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am. J. Respir. Crit. Care Med.* **180,** 388–395 (2009).
10. Innes, A. L. *et al.* Epithelial mucin stores are increased in the large airways of smokers with airflow obstruction. *Chest* **130,** 1102–1108 (2006).
11. Ordoñez, C. L. *et al.* Mild and moderate asthma is associated with airway goblet cell hyperplasia and abnormalities in mucin gene expression. *Am. J. Respir. Crit. Care Med.* **163,** 517–523 (2001).

12. Kawakubo, M. *et al.* Natural antibiotic function of a human gastric mucin against helicobacter pylori infection. *Science* **305,** 1003–1006 (2004).
13. Johansson, M. E. *et al.* The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proc. Natl Acad. Sci. USA* **105,** 15064–15069 (2008).
14. Van der Sluis, M. *et al. Muc2*-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* **131,** 117–129 (2006).
15. Velcich, A. *et al.* Colorectal cancer in mice genetically deficient in the mucin *Muc2*. *Science* **295,** 1726–1729 (2002).
16. Hasnain, S. Z. *et al.* Muc5ac: a critical component mediating the rejection of enteric nematodes. *J. Exp. Med.* **208,** 893–900 (2011).
17. Preciado, D. *et al.* Muc5b is the predominant mucin glycoprotein in chronic otitis media fluid. *Pediatr. Res.* **68,** 231–236 (2010).
18. Labandeira-Rey, M. *et al. Staphylococcus aureus* panton-valentine leukocidin causes necrotizing pneumonia. *Science* **315,** 1130–1133 (2007).
19. Stark, M. A. *et al.* Phagocytosis of apoptotic neutrophils regulates granulopoiesis via IL-23 and IL-17. *Immunity* **22,** 285–294 (2005).
20. Patnode, M. L. *et al.* Galactose-6-o-sulfotransferases are not required for the generation of siglec-F ligands in leukocytes or lung tissue. *J. Biol. Chem.* **288,** 26533–26545 (2013).
21. Moser, C. *et al.* β-defensin 1 contributes to pulmonary innate immunity in mice. *Infect. Immun.* **70,** 3068–3072 (2002).
22. Manitz, M. P. *et al.* Loss of *S100a9* (*Mrp14*) results in reduced interleukin-8-induced CD11b surface expression, a polarized microfilament system, and diminished responsiveness to chemoattractants *in vitro*. *Mol. Cell. Biol.* **23,** 1034–1043 (2003).
23. Flo, T. H. *et al.* Lipocalin 2 mediates an innate immune response to bacterial infection by sequestrating iron. *Nature* **432,** 917–921 (2004).
24. Ward, P. P., Mendoza-Meneses, M., Cunningham, G. A. & Conneely, O. M. Iron status in mice carrying a targeted disruption of lactoferrin. *Mol. Cell. Biol.* **23,** 178–185 (2003).
25. O'Riordan, T. G., Zwang, J. & Smaldone, G. C. Mucociliary clearance in adult asthma. *Am. Rev. Respir. Dis.* **146,** 598–603 (1992).
26. Goodman, R. M., Yergin, B. M., Landa, J. F., Golivanux, M. H. & Sackner, M. A. Relationship of smoking history and pulmonary function tests to tracheal mucous velocity in nonsmokers, young smokers, ex-smokers, and patients with chronic bronchitis. *Am. Rev. Respir. Dis.* **117,** 205–214 (1978).
27. Rowe, S. M., Miller, S. & Sorscher, E. J. Cystic fibrosis. *N. Engl. J. Med.* **352,** 1992–2001 (2005).
28. Talbot, T. R. *et al.* Asthma as a risk factor for invasive pneumococcal disease. *N. Engl. J. Med.* **352,** 2082–2090 (2005).
29. Lee, T. A., Weaver, F. M. & Weiss, K. B. Impact of pneumococcal vaccination on pneumonia rates in patients with COPD and asthma. *J. Gen. Intern. Med.* **22,** 62–67 (2007).
30. Hendley, J. O. Clinical practice. Otitis media. *N. Engl. J. Med.* **347,** 1169–1174 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank F. Ttofali, D. Harper, D. Raclawska, V. Mdoe, C. Ramsey and J. Parker-Thornburg for their assistance. We also thank K. Naff and the MD Anderson Cancer Center Department of Veterinary Medicine and Surgery for support in animal care. This work was supported by the National Institutes of Health Grants R01 HL080396 (C.M.E.); R01 AA008769 (J.H.S.), R01 HL109517 (W.J.J.); R01 HL114381 (P.M.H.); R01 HL097000 (B.F.D.), P01 HL108808, P01 HL110873, P50 HL107168, P30 DK065988 (R.C.B.), Medical Research Council Grant G1000450 (D.J.T.) and Cystic Fibrosis Foundation Grants 06I0 (C.M.E.) and RDP R026-CR11 (R.C.B.). Additional support was provided by National Institutes of Health Cancer Center Support Grants CA016672 for the MD Anderson Cancer Center and CA046934 for the University of Colorado transgenic mouse facilities; by CA016086 for the UNC Biomedical Research Imaging Center Small Animal Imaging Facility, and for the UNC Michael Hooker Microscopy Facility funded by an anonymous private donor.

Author Contributions M.G.R., A.L.-B. and A.A.F designed and performed survival, histological, particle clearance, inflammation, and infectious agent identification experiments, performed and collected data for *S. aureus* infection, macrophage and neutrophil identification experiments, and cytokine analyses. A.M.W., R.C.K., C.E.H. and D.A.S. generated and assisted in studies in *Scgb1a1-Muc5b* mice. M.M.M., R.M.B, I.R., A.S.B., M.G.B., W.K.O., K.A.T., S.C.F., A.C.-G., J.M.D. and C.A.L. performed infectious pneumonia and infectious agent identification experiments. S.N.A., L.K.B., A.S.S. and Y.M.P. constructed and generated *Muc5b* knockout mice. S.E.E., M.M.D.G., S.J.M. and M.J.T. assisted in the design and performance of inflammation studies. B.R.G., R.A., H.K.-Q. and M.R.B. assisted in the design and performance of hypoxemia studies. J.H.S., S.M.D. and B.R.G. assisted in the design and performance of tracheal and nasal mucociliary function studies. D.J.T. and K.R. provided purified MUC5B protein. W.J.J. and L.B. designed and performed macrophage activation and apoptosis assays. I.V.Y, P.M.H., P.G.W., C.W.D., R.C.B and B.F.D. assisted in the analysis and interpretation of data, and C.W.D. provided Muc5b antisera. C.M.E. designed the study, analysed data, and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.M.E. (Christopher.Evans@ucdenver.edu).

LETTER

doi:10.1038/nature12791

Cytokinin signalling inhibitory fields provide robustness to phyllotaxis

Fabrice Besnard[1]†, Yassin Refahi[2], Valérie Morin[1]†*, Benjamin Marteaux[1]*, Géraldine Brunoud[1]*, Pierre Chambrier[1], Frédérique Rozier[1], Vincent Mirabet[1,3], Jonathan Legrand[1,2,3], Stéphanie Lainé[1], Emmanuel Thévenon[4], Etienne Farcot[2]†, Coralie Cellier[1], Pradeep Das[1,3], Anthony Bishopp[5]†, Renaud Dumas[4], François Parcy[4], Ykä Helariutta[5], Arezki Boudaoud[1,3], Christophe Godin[2], Jan Traas[1], Yann Guédon[2] & Teva Vernoux[1]

How biological systems generate reproducible patterns with high precision is a central question in science[1]. The shoot apical meristem (SAM), a specialized tissue producing plant aerial organs, is a developmental system of choice to address this question. Organs are periodically initiated at the SAM at specific spatial positions and this spatiotemporal pattern defines phyllotaxis. Accumulation of the plant hormone auxin triggers organ initiation[2–5], whereas auxin depletion around organs generates inhibitory fields that are thought to be sufficient to maintain these patterns and their dynamics[4,6–13]. Here we show that another type of hormone-based inhibitory fields, generated directly downstream of auxin by intercellular movement of the cytokinin signalling inhibitor *ARABIDOPSIS* HISTIDINE PHOSPHOTRANSFER PROTEIN 6 (AHP6)[14], is involved in regulating phyllotactic patterns. We demonstrate that AHP6-based fields establish patterns of cytokinin signalling in the meristem that contribute to the robustness of phyllotaxis by imposing a temporal sequence on organ initiation. Our findings indicate that not one but two distinct hormone-based fields may be required for achieving temporal precision during formation of reiterative structures at the SAM, thus indicating an original mechanism for providing robustness to a dynamic developmental system.

In the most widely accepted theory for phyllotaxis, spatiotemporal patterns of organ initiation at the SAM result from the combined effect of inhibitory fields produced by existing organs. As these organs are displaced away from the SAM by growth, new organs form sequentially at positions where the sum of the inhibitory effects is the lowest[6–8]. Both the position of new organs and the time delay between organ initiations (or plastochron) are emergent parameters of this dynamical system[8]. Strong evidence indicates that, notably through a network of polarly localized PIN-FORMED 1 (PIN1) efflux carriers, polar auxin transport not only controls accumulation of auxin at the site of organ initiation but also creates inhibitory fields around organs by auxin depletion[4,9–13]. It has therefore been proposed that the auxin transport system could be sufficient to control both the spatial and the temporal dynamics of phyllotaxis[4].

Here we re-evaluated this proposition by considering the role of cytokinin during organ initiation in *Arabidopsis thaliana*. Cytokinin regulates the size of the stem cell niche (and thus of the SAM) and this can effect phyllotaxis by modifying the geometry of the SAM[15,16]. To explore a possible role of cytokinin directly in organ initiation, we re-analysed recent transcriptomic data for different domains of the SAM[17] and identified *AHP6* (ref. 14) as the only candidate gene encoding a cytokinin signalling effector specifically enriched in organs (Fig. 1a, Supplementary Figs 1 and 2a). Using *in situ* hybridization, we confirmed that *AHP6* is specifically expressed during organ initiation and development (Fig. 1b, c, Supplementary Fig. 2b–h)[18]. Wild-type *Arabidopsis* plants display spiral phyllotaxis, resulting in consecutive organs generally distributed on the stem at a divergence angle close to a 137.5° canonical angle (noted as α; Fig. 1d). In contrast, both *ahp6-1* and *ahp6-3* null mutants as well as the *ahp6-1*/*ahp6-3* (*ahp6-1/3*) trans-heterozygote showed obvious modifications in organ arrangements along the stem (Fig. 1e–g). We also observed supernumerary petals and sepals in flowers (Supplementary Fig. 3), indicating that *AHP6* regulates phyllotaxis throughout inflorescence development.

To further characterize the stem architecture of *ahp6* mutants, we analysed sequences of successive divergence angles between organs on the inflorescence stems from a large population of *ahp6-1* and wild-type plants. This analysis demonstrated a notable amount of non-canonical divergence angles in wild-type plants and a large increase in the occurrence of such angles in most mutants (Supplementary Fig. 4a, b). Notably, an 'M-shaped' motif corresponding approximately to the angle sequence 2α, $360-\alpha$, 2α appeared much more frequently in *ahp6* than in wild-type sequences (Fig. 1h, i and Supplementary Fig. 4c–f). It was not associated with changes in the structure of the stem, such as twisting, that could modify angles between siliques (Supplementary Fig. 5). This motif can theoretically arise if two consecutive organs in a canonical sequence are permuted along the stem (Fig. 1j)[19]. By applying a stochastic and a combinatorial model to analyse the divergence angle sequences (see Methods), we showed that over 95% of the non-canonical angles can indeed be explained by permutations of the insertion order of 2 to 3 organs in both wild-type and *ahp6* plants (Fig. 1k, l and Supplementary Fig. 4c–f). We further demonstrated an increase by 2.4-fold and 17.6-fold of permutations involving 2 and 3 organs, respectively, in *ahp6* mutants compared to wild-type plants (Fig. 1k). Altogether, our data indicate that *AHP6* is required for buffering an intrinsic instability of phyllotaxis leading to permutations in the order of organ insertions along the stem.

We next used scanning electron microscopy (SEM) to study the geometry of the *ahp6* mutant SAM. In wild-type meristems the spatial organization of the organs usually followed the expected phyllotaxis and consecutive organs showed clear differences in size (Fig. 2a). In contrast, pairs or triplets of young organs at quasi-identical developmental stages occurred in most *ahp6* meristems (Fig. 2b). This observation was confirmed using a *LEAFY* (*LFY*) promoter driving GFP (*pLFY::GFP*) flower-specific marker line (Fig. 2c, d), demonstrating that the loss of *AHP6* leads to simultaneous development of flowers instead of the mostly sequential outgrowth observed in wild-type meristems. As the size of the stem-cell niche and of the meristem were not significantly affected by the *ahp6* mutation (Supplementary Fig. 6), these data indicate that *AHP6* is required for regulating the sequence of organ

[1]Laboratoire de Reproduction et Développement des Plantes, CNRS, INRA, ENS Lyon, UCBL, Université de Lyon, 69364 Lyon, France. [2]Virtual Plants INRIA/CIRAD/INRA Project Team, UMR AGAP, Institut de Biologie Computationelle, 34095 Montpellier, France. [3]Laboratoire Joliot-Curie, CNRS, ENS Lyon, Université de Lyon, 69364 Lyon, France. [4]Laboratoire Physiologie Cellulaire et Végétale, CEA, CNRS, INRA, UJF, 38041 Grenoble, France. [5]Institute of Biotechnology/Department of Biosciences, University of Helsinki, FIN-00014, Finland. †Present addresses: IBENS, ENS, 75005 Paris, France (F.B.); UMR CNRS 5534, Université Claude Bernard Lyon I, Bâtiment Gregor Mendel, 16 rue Raphaël Dubois, 69622 Villeurbanne, France (V.M.); University of Nottingham, University Park, Nottingham NG7 2RD, UK (E.F); University of Nottingham, Sutton Bonington LE12 5RD, UK (A.Bi.).
*These authors contributed equally to this work.

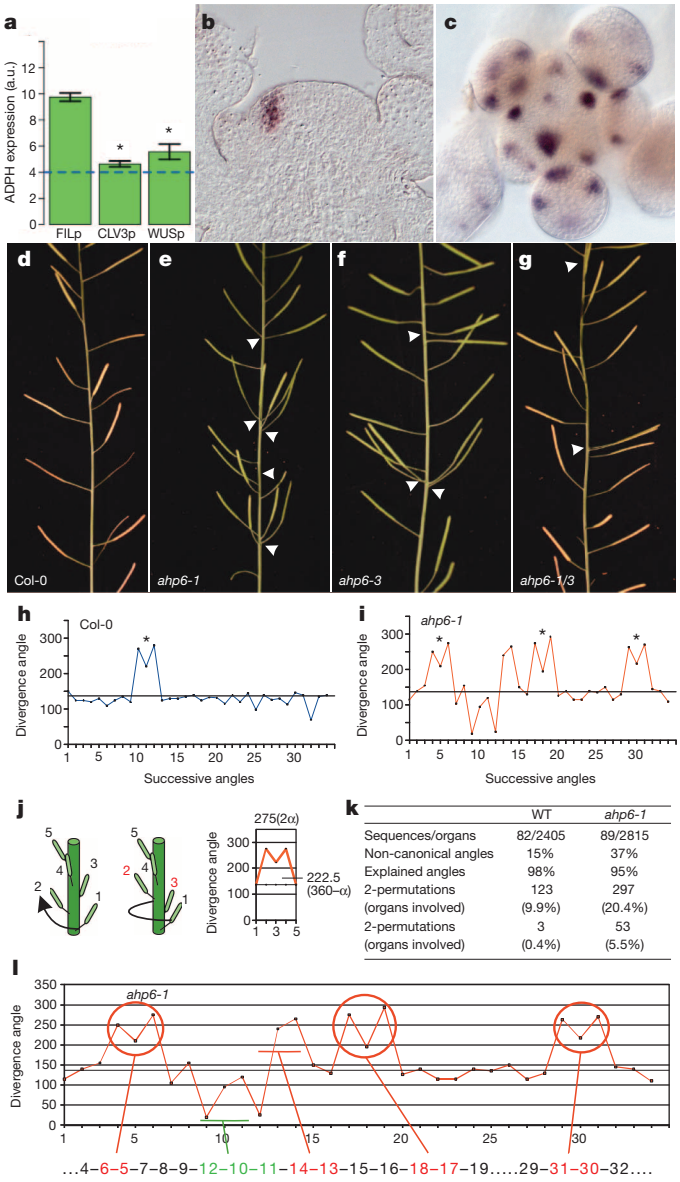


Figure 1 | *AHP6* is expressed in lateral organs and regulates inflorescence architecture. **a**, *AHP6* (AT1G80100) expression in genome-wide data for lateral organs (FILp), stem cells (CLV3p) and the WUSCHEL domain (WUSp)[17]. An asterisk indicates statistically different from FILp (Methods). Dashed line indicates non-specific background level. Expression measured in arbitrary units (a.u.). **b**, **c**, *AHP6 in situ* hybridization on section (**b**) and whole-mount tissues (**c**). **d**–**g**, Wild-type (**d,** Col-0) and *ahp6* (**e**–**g**) inflorescences. Arrowheads (**d**–**g**) indicate perturbations. **h**, **i**, Representative angle sequences for wild type (**h**, $n = 82$) and *ahp6-1* (**i**, $n = 89$). Asterisk indicates M-shaped motif. **j**, Explaining the M-shaped motif: canonical insertion order (left), the one obtained by permutating two organs (centre) and angle sequences (right). **k**, Quantification of permutations. Wild type (WT). **l**, Permutations in the *ahp6-1* sequence from (**i**).

initiation and/or organ growth following initiation to restrict concomitant development of organs.

To discriminate between these possible scenarios, we followed the expression of a nuclear-localized *DR5::VENUS*[5] over several days in *ahp6* and wild-type meristems. The synthetic auxin-inducible *DR5* reporter allows monitoring of the recruitment of organ founder cells starting from initium i1 and in primordia from early P1 onwards (Supplementary Fig. 7)[5,9,11,20]. Live imaging of *DR5::VENUS* demonstrated that, although the mean rate of organ initiation (the mean plastochron) is comparable in wild-type and *ahp6* (Supplementary Fig. 8), the loss of *AHP6* results in a strong increase in concomitant organ initiation and, at a very low frequency (8 out of 255 initiation events), in permutations in the order of organ initiation (Fig. 2e–g and Supplementary Fig. 9). On the contrary, the relative angle between organs and the position of their initiation relative to the centre of the meristem were unaffected by the *ahp6* mutation (Fig. 2h, i). In addition, co-initiated organs were also generated at a similar radial distance from the centre independently of the genotype and with a standard deviation smaller than one cell diameter (Supplementary Fig. 10a)[21]. The rate of organ displacement away from the centre of the SAM was also similar between *ahp6* and wild-type, indicating that AHP6 does not influence growth (Supplementary Fig. 10b). Thus, although organs are initiated at a precise radial distance and with a precise divergence angle, frequent organ co-initiations are observed, indicating an irregular plastochron. *AHP6* is then required to stabilize the plastochron by limiting organ co-initiations. Our data further indicate that organ permutations on the stem result mostly from organ co-initiations. Also, the frequency of perturbations in the sequence of organ initiation at the SAM is higher than the frequency of permutations on the stem in both wild-type and *ahp6* (28% and 47% compared to 10% and 25% for wild-type and *ahp6,* respectively). This indicates that co-initiated organs are sorted when the internode is established either in a normal or in an inverted order, only the latter resulting in permutations on the inflorescence stem.

As auxin activates directly *AHP6* transcription in root tissues[22], we next investigated whether the spatiotemporal pattern of *AHP6* transcription in the SAM could be controlled by auxin. Co-visualization of a *pAHP6::GFP* transcriptional reporter that recapitulates *AHP6* expression pattern (Fig. 3a) and *DR5::VENUS* showed that *AHP6* is activated one plastochron after *DR5* activation (Fig. 3b–d and Supplementary Video 1). *AHP6* expression in the SAM was also lost in *monopteros* (*mp*), a mutant in a major transcriptional effector of auxin signalling in the SAM (Fig. 3e)[23]. In addition, using electrophoretic mobility shift assays (EMSAs), we mapped binding of MONOPTEROS to three out of six locations of putative ARF binding sites in the *AHP6* promoter (Fig. 3f). This indicates that *AHP6* is activated downstream of auxin by MONOPTEROS.

The temporal delay between *DR5* and *AHP6* activation, together with the fact that *AHP6* is required for the earliest steps of organ initiation, indicates that *AHP6* acts non-cell-autonomously on the temporal sequence of organ initiation. We thus monitored the distribution in the SAM of a functional AHP6–GFP protein fusion expressed under the endogenous *AHP6* promoter[14]. We observed sharp gradients of AHP6–GFP centred on primordia and extending beyond their boundaries, indicating that intercellular movement of the protein creates fields of AHP6 around organs (Fig. 3g). AHP6–GFP fluorescence quantification further demonstrated that AHP6 movement creates a non-cell-autonomous differential in AHP6 levels between the predicted i1 and i2 sites (Fig. 3h, i; Supplementary Fig. 11), AHP6 levels being 1.47-fold higher (± 0.32 with $n = 12$ meristems; *t*-test: one-sided $P = 2 \times 10^{-4}$) at the i2 site. Fluorescence profiles taken through i1 or i2 from the two closest primordia (P3 and P5 or P2 and P4, respectively) further indicate that the proximity of P2 and to a lesser extent P4 allows for the higher AHP6 level at the i2 site (Fig. 3j and Supplementary Fig. 12). These profiles also allowed visualization of changes in AHP6 levels produced by primordia; AHP6 levels increase first and then decrease strongly from P4 onwards (Fig. 3h, j). AHP6 levels at i1 are then lower owing not only to an increased distance between i1 and P3 (P2 being further away), but also to lower AHP6 levels produced by P5. Thus, the differential in AHP6 levels between the i1 and i2 sites, that could also be visualized in plants co-expressing AHP6–GFP and *DR5::VENUS* (Fig. 3k), results from both the geometry of the SAM and dynamic changes in AHP6 levels during flower primordia development. AHP6 movement could be blocked by fusing AHP6 to a triple-VENUS (3×VENUS; Fig. 3l and Supplementary Fig. 13a–c). The AHP6–3×VENUS protein expressed under the endogenous *AHP6* promoter (*pAHP6::AHP6–3×VENUS*) was still functional because it could complement the cell-autonomous loss of protoxylem phenotype in *ahp6* roots (Supplementary Fig. 13d–g)[14]. However,

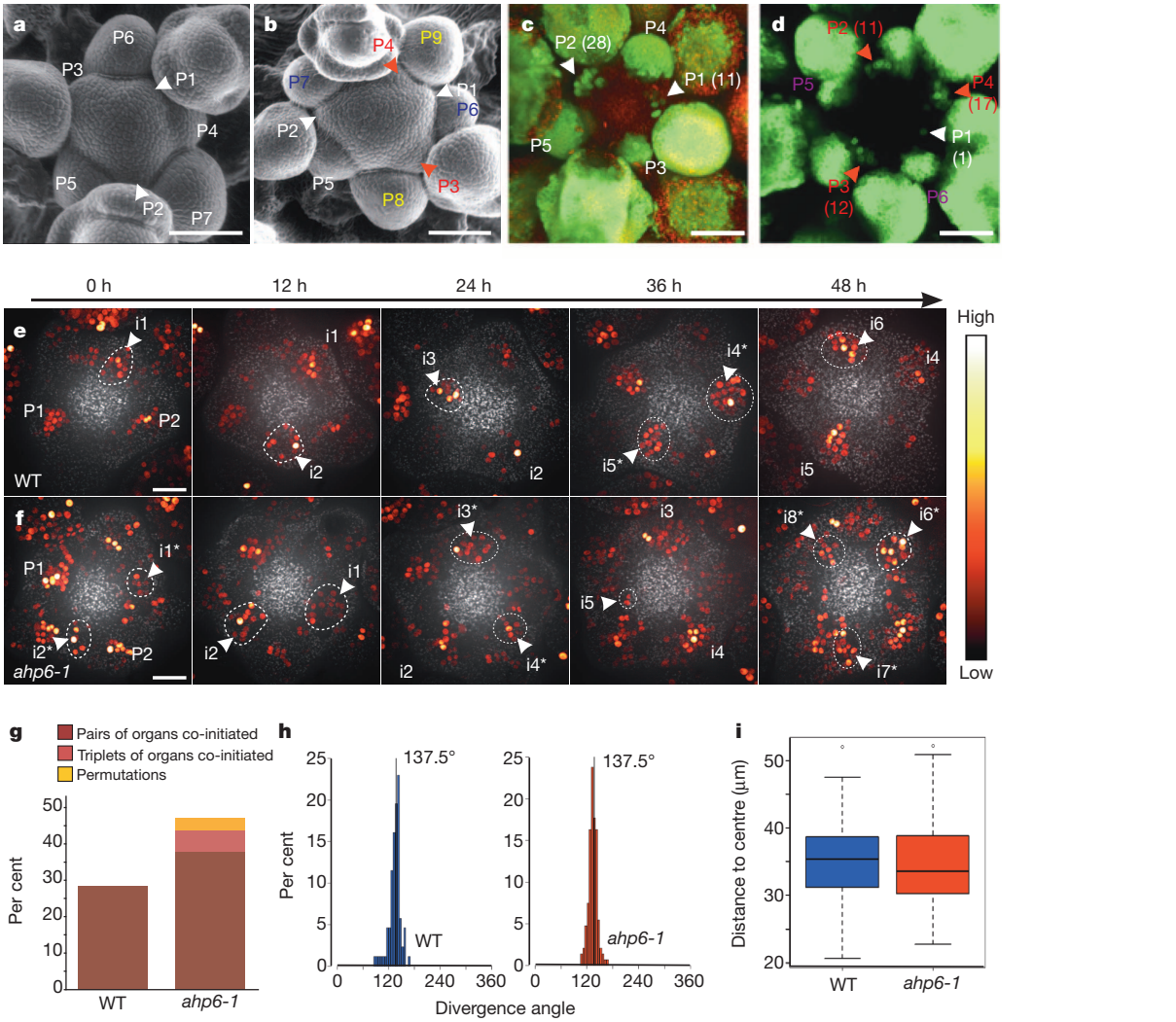


Figure 2 | *AHP6* regulates the plastochron. **a**, **b**, SEM of wild type (**a**, $n = 15$) and *ahp6-1* (**b**, $n = 17$). Letters of identical colours (except white) indicate morphologically identical stages. **c**, **d**, *pLFY::GFP* expression in wild type (**c**) and *ahp6-1* (**d**). Numbers within the brackets indicate number of cells. **e**–**i**, Organ initiation timing and position in wild type ($n = 20$) and *ahp6-1* ($n = 33$) expressing *DR5::VENUS*. Representative time courses (**e**, wild type; **f**, *ahp6-1*); organ co-initiations/permutations (**g**); relative angles (**h**); and radial position of initiation (**i**). Primordia (P) numbered from youngest to oldest, initia (**i**) from oldest to youngest. Asterisks indicate co-initiated organs. Autofluoresence visible in red (**c**, **d**) or grey channels (**e**, **f**). Scale bars, 50 µm (**a**–**d**) or 20 µm (**e**, **f**).

in contrast to *pAHP6::AHP6–VENUS*, *pAHP6::AHP6–3×VENUS* could not complement the *ahp6-1* inflorescence phenotype (Fig. 3m). This suggests that AHP6 movement in the SAM generates a differential in AHP6 levels between i1 and presumptive i2, which is required for sequential organ initiation. Note that a narrow region with a lower level of AHP6 was often observed in the vicinity of the expected i2 position, although the proximity of P2 and P4 limited its size compared to i1 (Fig. 3h, j and Supplementary Fig. 12). Thus, slight errors in i2 positioning can almost abolish the i1/i2 AHP6 differential. This could explain the occurrence of co-initiations even in wild-type meristems.

As *AHP6* functions as a cytokinin signalling inhibitor in the root[14], its spatial distribution in the SAM could create differential cytokinin signalling capacities between i1 and presumptive i2. Indeed, comparison of wild-type, *ahp6* mutants and *35S::AHP6* plants (Fig. 4a) showed that *AHP6* levels negatively modulate the induction by cytokinin of several primary cytokinin response genes[16,24,25] in shoot tissues. Expression of the cytokinin-inducible synthetic reporter *TCS::GFP*[25] was also extended in *ahp6* meristems (Fig. 4b, c), indicating that AHP6 negatively regulates the spatial distribution of cytokinin signalling in the SAM. In addition, we found that *TCS::GFP* and *DR5::VENUS* were activated together during organ initiation in partly overlapping domains in wild-type plants: *TCS::GFP* expression was absent from presumptive i2 but started to be expressed at low levels in i1 and increased in P1 onwards, thus demonstrating a progressive activation of cytokinin signalling (Fig. 4d, f–h, l). In *ahp6* meristems, *TCS::GFP* was on the contrary already expressed in i2 at a level similar to that observed in P1 in wild type and increased further compared with wild type from i1 onwards (Fig. 4e, i–k, l), indicating that AHP6 protein distribution regulates the spatiotemporal pattern of cytokinin signalling during organ initiation. Our results thus show that AHP6 distribution in the SAM creates a differential in cytokinin signalling between i1 (higher cytokinin signalling) and presumptive i2 (lower cytokinin signalling).

The negative correlation between AHP6 protein levels and *DR5* levels in i1 and presumptive i2 suggests that the differential in cytokinin signalling generated by AHP6 could act by modulating PIN-regulated auxin transport, as observed in root tissues[22,26,27]. However, we could not detect any changes in PIN1 levels or intracellular localization or in PIN1 polarity distribution in the SAM of *ahp6* mutants (Supplementary Fig. 14). This indicates that AHP6 is unlikely to affect auxin transport in the SAM but rather regulates organ initiation after organ positioning by auxin, by acting either in parallel with or downstream of auxin. Consistent with this hypothesis, the downregulation of several cytokinin

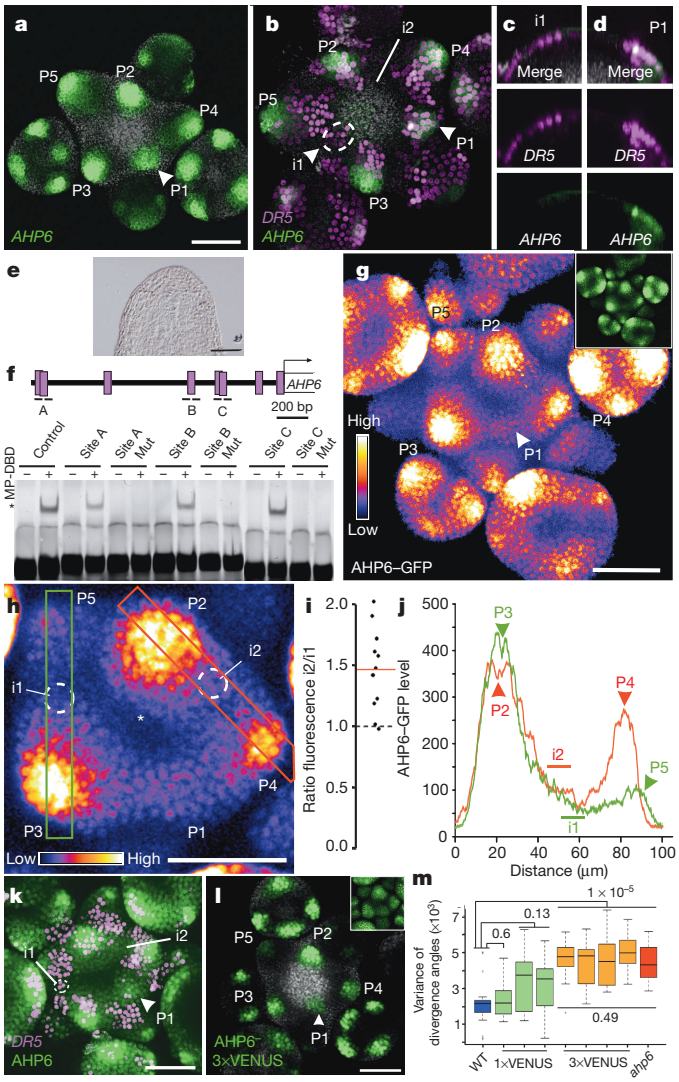


Figure 3 | AHP6 acts non-cell-autonomously downstream of auxin. **a**, *pAHP6::GFP*. **b**, *pAHP6::GFP* (green) and *DR5::VENUS* (magenta; $n = 6$). **c**, **d**, Longitudinal optical sections at i1 (**c**) and P1 (**d**). **e**, *AHP6* expression in *mp-S319*. **f**, Electrophoretic mobility shift assay (EMSA) using MONOPTEROS DNA-binding domain (MP-DBD). Cartoon shows putative ARF binding sites in *AHP6* promoter (1.6 kb). **g**–**k**, AHP6 protein distribution. **g**, *pAHP6::AHP6–GFP* ($n = 12$; inset shows original image). **h**, Close-up of (**g**). **i**, i2/i1 AHP6–GFP ratios. **j**, Fluorescence distribution along areas shown in (**h**). **k**, *pAHP6::AHP6–GFP* (green) and *DR5::VENUS* (magenta). **l**, *pAHP6::AHP6–3×VENUS* (inset: *pAHP6::AHP6–VENUS*). **m**, Boxplots of variance per individual of divergence angles between siliques in wild type ($n = 19$), *ahp6-1 pAHP6::AHP6–VENUS* ($n = 23$, 19 and 17, respectively), *ahp6-1 pAHP6::AHP6–3×VENUS* ($n = 17$, 17, 18 and 20, respectively) and *ahp6-1* ($n = 18$). *P* values: two-sided Kruskal–Wallis. Autofluorescence visible in grey (**a**–**d**, **l**). Scale bars, 50 μm.

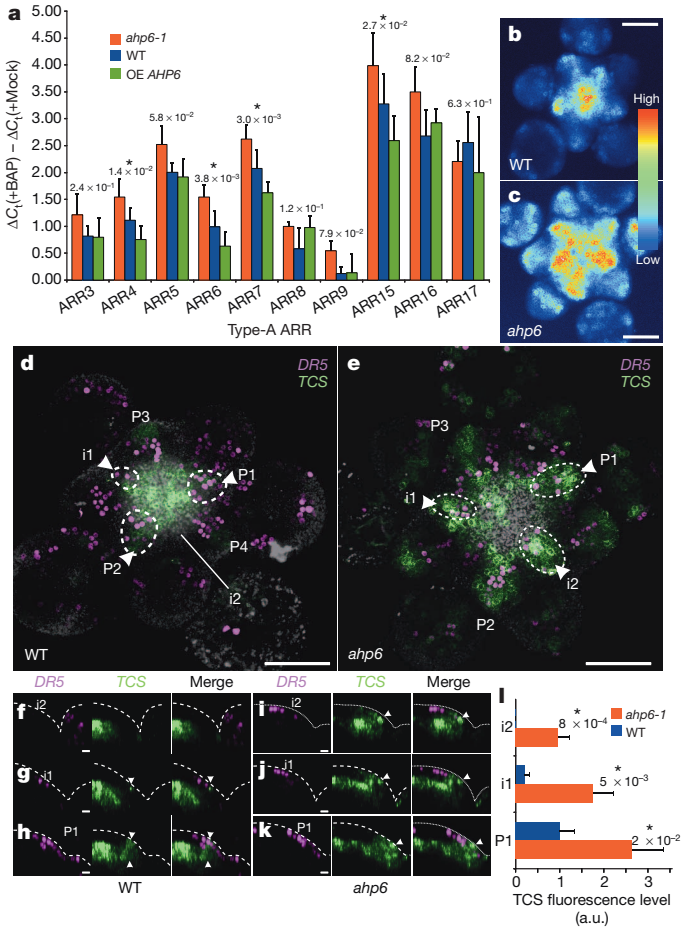


Figure 4 | AHP6 inhibitory fields generate patterns of cytokinin signalling in the meristem. **a**, qRT–PCR of *ARABIDOPSIS RESPONSE REGULATORS* (*ARRs*) induction by cytokinin (100 nM BAP) in wild-type, *ahp6-1* and *35S::AHP6* (overexpression, OE). *P* value shows two-sided ANOVA ($n = 3$), $*P = 0.05$. Error bars show s.d. **b**, **c**, *TCS::GFP* in wild-type (**b**; WT: $n = 10$) and *ahp6-1* (**c**; $n = 9$). **d**–**k**, *DR5::VENUS* (magenta) and *TCS::GFP* (green) in wild type (**d**, **f**–**h**) and *ahp6-1* (**e**, **i**–**k**). Longitudinal optical section with both merged and separate channels in i2, i1 and P1 are shown (WT: **f**–**h**; *ahp6-1*: **i**–**k**). Arrowheads indicate *TCS::GFP* in primordia; **l**, TCS::GFP quantification. *P* values from one-sided Mann–Whitney *U*-test (WT: $n = 5$; *ahp6-1*: $n = 8$; $*P = 0.05$). Error bars represent s.d. Grey autofluorescence visible (**d**, **e**). Scale bars, 50 μm (**b**–**e**), 10 μm (**f**–**k**).

signalling inhibitors can partially restore organ initiation in the auxin signalling deficient *mp* mutant[28]. A plausible scenario would then be that lower levels of AHP6 in i1 compared to presumptive i2 promote i1 initiation. Inversely, higher concentration of AHP6 in i2 would repress cytokinin signalling and organ initiation, allowing for a time delay between i1 and i2 initiations.

In conclusion, our results indicate that, although the spatial position of new organs at the SAM is robustly determined by the auxin-based inhibitory fields[4,5,9–13], the dynamics of these fields leads to a noisy plastochron. Accordingly, a recent theoretical study demonstrated that noise induces principally irregularities of the plastochron in a phyllotactic model[29]. Our results further indicate that the noise on the plastochron is then filtered out in part by non-canonical inhibitory fields generated by AHP6 movement downstream of the primary auxin fields. In this scheme, plastochron robustness would be increased by conversion of the spatial information provided by cytokinin signalling inhibitory fields into a roughly periodic temporal sequence of auxin-induced organogenesis (Supplementary Fig. 15).

METHODS SUMMARY

The *Arabidopsis thaliana* Columbia (Col-0) ecotype was used except for *pWUS::GFP* and *pLFY::GFP* (Ws). *TCS::GFP* was generated by transformation of a previously described plasmid[25]. *pAHP6::AHP6–VENUS/3×VENUS* and *35S::AHP6* were constructed by Gateway recombination, using a 2,494-bp AHP6 genomic fragment and full-length cDNA respectively, and introduced in plants.

RNA *in situ* hybridization and SEM were performed as described[2,17], with minor modifications for the whole-mount RNA *in situ* hybridization. To analyse xylem defects, roots were cleared with chloral hydrate. Light microscopy images were obtained using either transmission or laser-scanning confocal microscopes. Culture and live-imaging of SAM were performed as previously described[30], except for adding 555 nM BAP to the culture medium. Images were processed and analysed using Image J (http://rsbweb.nih.gov/ij/). The PIN1 network was analysed from PIN1 whole-mount immunolocalization as described[12] and by computing (under

Python) a PIN1 polarity coherence index for each cell to estimate the local coherence of PIN1 orientations.

Real-time polymerase chain reaction (RT–PCR) was performed using the SYBR green reagent kit (Roche) on shoots of 7-day-old seedlings as described[13]. Microarray data were analysed under R using GC robust multi-array average (gcRMA) to estimate expression, LIMMA for statistical tests and *Q*-values to correct for multiple testing. For other data, statistical analyses were done using *t*-tests (or ANOVA for multiple comparisons) only when normal distributions and homoscedasticity of the response variables were verified, and non-parametrical Mann–Whitney tests used (or Kruskal–Wallis tests for multiple comparisons) otherwise.

EMSA was performed using 30–40-bp double-stranded oligonucleotides labelled with Cy5 and corresponding to the *AHP6* promoter sequence with or without mutations of putative ARF binding sites.

To identify permutations in the sequences of divergence angle measured on inflorescences, a combinatorial mixture model and a hidden Markov chain model were used (see Methods).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 March; accepted 17 October 2013.

Published online 15 December 2013.

1. Lander, A. D. Pattern, growth, and control. *Cell* **144,** 955–969 (2011).
2. Vernoux, T., Kronenberger, J., Grandjean, O., Laufs, P. & Traas, J. PIN-FORMED 1 regulates cell fate at the periphery of the shoot apical meristem. *Development* **127,** 5157–5165 (2000).
3. Reinhardt, D., Mandel, T. & Kuhlemeier, C. Auxin regulates the initiation and radial position of plant lateral organs. *Plant Cell* **12,** 507–518 (2000).
4. Reinhardt, D. *et al.* Regulation of phyllotaxis by polar auxin transport. *Nature* **426,** 255–260 (2003).
5. Heisler, M. G. *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem. *Curr. Biol.* **15,** 1899–1911 (2005).
6. Mitchison, G. J. Phyllotaxis and the Fibonacci series. *Science* **196,** 270–275 (1977).
7. Veen, A. H. & Lindenmayer, A. Diffusion mechanism for phyllotaxis: theoretical physico-chemical and computer study. *Plant Physiol.* **60,** 127–139 (1977).
8. Douady, S. & Couder, Y. Phyllotaxis as a dynamical self organizing process. Part II: the spontaneous formation of a periodicity and the coexistence of spiral and whorled patterns. *J. Theor. Biol.* **178,** 275–294 (1996).
9. de Reuille, P. B. *et al.* Computer simulations reveal properties of the cell–cell signaling network at the shoot apex in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **103,** 1627–1632 (2006).
10. Jönsson, H., Heisler, M. G., Shapiro, B. E., Meyerowitz, E. M. & Mjolsness, E. An auxin-driven polarized transport model for phyllotaxis. *Proc. Natl Acad. Sci. USA* **103,** 1633–1638 (2006).
11. Smith, R. S. *et al.* A plausible model of phyllotaxis. *Proc. Natl Acad. Sci. USA* **103,** 1301–1306 (2006).
12. Stoma, S. *et al.* Flux-based transport enhancement as a plausible unifying mechanism for auxin transport in meristem development. *PLOS Comput. Biol.* **4,** e1000207 (2008).
13. Vernoux, T. *et al.* The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Mol. Syst. Biol.* **7,** 508 (2011).
14. Mähönen, A. P. Cytokinin signaling and its inhibitor AHP6 regulate cell fate during vascular development. *Science* **311,** 94–98 (2006).
15. Giulini, A., Wang, J. & Jackson, D. Control of phyllotaxy by the cytokinin-inducible response regulator homologue *ABPHYL1*. *Nature* **430,** 1031–1034 (2004).
16. Leibfried, A. *et al.* WUSCHEL controls meristem function by direct regulation of cytokinin-inducible response regulators. *Nature* **438,** 1172–1175 (2005).
17. Yadav, R. K., Girke, T., Pasala, S., Xie, M. & Reddy, G. V. Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche. *Proc. Natl Acad. Sci. USA* **106,** 4941–4946 (2009).
18. Bartrina, I., Otto, E., Strnad, M., Werner, T. & Schmülling, T. Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in *Arabidopsis thaliana*. *Plant Cell* **23,** 69–80 (2011).
19. Couder, Y. Initial transitions, order and disorder in phyllotactic patterns: the ontogeny of *Helanthus annuus*. A case study. *Acta Societates Botanicarum Poloniae* **67,** 129–150 (1998).
20. Benková, E. *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115,** 591–602 (2003).
21. Laufs, P., Grandjean, O., Jonak, C., Kiêu, K. & Traas, J. Cellular parameters of the shoot apical meristem in *Arabidopsis*. *Plant Cell* **10,** 1375–1390 (1998).
22. Bishopp, A. *et al.* A mutually inhibitory interaction between auxin and cytokinin specifies vascular pattern in roots. *Curr. Biol.* **21,** 917–926 (2011).
23. Hardtke, C. S. & Berleth, T. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J.* **17,** 1405–1411 (1998).
24. To, J. P. C. *et al.* Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin signaling. *Plant Cell* **16,** 658–671 (2004).
25. Müller, B. & Sheen, J. Cytokinin and auxin interaction in root stem-cell specification during early embryogenesis. *Nature* **453,** 1094–1097 (2008).
26. Dello Ioio, R. *et al.* A genetic framework for the control of cell division and differentiation in the root meristem. *Science* **322,** 1380–1384 (2008).
27. Marhavý, P. *et al.* Cytokinin modulates endocytic trafficking of PIN1 auxin efflux carrier to control plant organogenesis. *Dev. Cell* **21,** 796–804 (2011).
28. Zhao, Z. *et al.* Hormonal control of the shoot stem-cell niche. *Nature* **465,** 1089–1092 (2010).
29. Mirabet, V., Besnard, F., Vernoux, T. & Boudaoud, A. Noise and robustness in phyllotaxis. *PLoS Comput. Biol.* **8,** e1002389 (2012).
30. Fernandez, R. *et al.* Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution. *Nature Methods* **7,** 547–553 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Heisler, B. Müller and D. Weijers for sharing materials; A. Miyawaki for VENUS; D. Mast for help with plant analysis; C. Chamot and C. Lionnet (PLATIM) for help with confocal microscopy and ImageJ (C. Chamot); S. Chamot for help with live imaging; C. Gauthier and A. Laugraud (PRABI, Université Lyon I) for help with statistical analyses; Xavier Jaurand (Pi2) for help with the SEM; Y. Couder, S. Douady, M. Bennett and M.-A. Félix for their insights and support; Y. Jaillais, O. Hamant and C. Scutt for comments on the manuscript. T.V. was supported by HFSPO CDA 0047/2007 (Human Frontier Science Program Organization) and ANR-07-JCJC-0115 (Agence National de la Recherche) grants; R.D., F.P. and T.V. by the ANR-12-BSV6-0005 grant (AuxiFlo); J.L., J.T., C.G., Y.H. and T.V. by a transnational EraSysBio Grant (iSAM); F.B. by a predoctoral grant of the French Ministry of Research; and Y.R. by a CJS grant from INRA.

Author Contributions F.B. and T.V. conceived and designed the experiments. F.B., V.Mo., B.M., G.B., P.C., F.R., J.L., S.L., C.C., E.T., P.D. and T.V. performed the experiments. Y.R., E.F., C.G., F.B., G.B., T.V. and Y.G. performed the mathematical analysis of phyllotaxis. V.Mi., G.B. and A.Bo. analysed the auxin transport network. R.D., F.P., A.Bi., Y.H. and J.T. provided reagents/materials. F.B. and T.V. analysed the data with inputs from all the authors. F.B. and T.V. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.V. (teva.vernoux@ens-lyon.fr).

METHODS

Plant material and growth conditions. The *ahp6-1*, *ahp6-3*, *mp-S319*, *pLFY::GFP*, *pWUS::GFP*, *pAHP6::GFP*, *pPIN1::PIN1–GFP*, *pAHP6::AHP6–GFP*, *pARR15::GFP*, *DR5::GFP* and *DR5::VENUS* lines have been described[13,14,20,25,31–33] and are all in Col-0 except *pLFY::GFP* and *pWUS::GFP* (Ws). All transgenic plants (see below) were generated in the Col-0 ecotype. Plants were grown *in vitro* on Murashige and Skoog (MS) medium or on soil at 20 °C in short-day conditions (8 h light/16 h darkness) for 4 to 5 weeks to synchronize them before being transferred in long-day conditions (16 h light/8 h darkness). Plants showing obvious developmental defects owing to adverse growth conditions were systematically discarded. Plants for live-imaging or for the analysis of phyllotactic sequences were always grown together at the same place within the growth chamber and with randomized position within each tray in order to minimize the effect of environmental fluctuations.

Cloning and constructs. A previously described plasmid[25] was introduced into plants by floral dipping[34] to generate the *TCS::GFP* line. To generate *35S::AHP6*, an *AHP6* full-length complimentary DNA (cDNA) (encompassing a 359-bp 5′ UTR upstream of the ATG and an 83-bp 3′ UTR after the stop codon) obtained by 5′ RACE–PCR was recombined into the Gateway binary vector dpGreenKanT under the control of the Cauliflower Mosaic Virus (CaMV) 35S promoter. dpGreenKanT was obtained by inserting a Nos terminator after the attR3 recombination site in Kanamycin-resistant version of the Gateway-compatible pGreen 0229 plasmid[35,36]. The *pAHP6::AHP6–3×VENUS* construct was generated by recombining in phase a 2,429-bp *AHP6* genomic fragment (from 1,594 bp upstream of the ATG to the stop codon), a 2×VENUS without stop codon and VENUS with stop codon into a gateway binary vector pK7m34GW (ref. 37). The plasmids were then transformed into Col-0 and *ahp6-1* plants, respectively, by floral dipping.

Real-time RT–PCR and microarray data analysis. The real-time RT–PCR analysis was designed to comply with standards of qRT–PCR[38,39] and performed on a StepOne Plus cycler (Applied Biosystems) using the SYBR green reagent kit (Roche). The 7-day-old plants grown *in vitro* were transferred for 24 h on plates containing either mock or 100 nM BAP before extraction of mRNA specifically from the shoot tissues using the Spectrum Plant total RNA kit (Sigma). The primers used are listed in Supplementary Table 1. The *TCTP* gene was used as a standard (Supplementary Table 2) and validated using BestKeeper[40]. PCR efficiency was calculated for each primer pairs using classical calibration dilution curve and used for the estimation of the ΔCt. Three biological replicates were tested and reactions were carried out using three technical replicates. The effect of the treatment was calculated as the difference between the ΔCt in the BAP-treated samples and the mock-treated samples (ΔΔCt). Statistical significance of the results was tested using an analysis of variance (ANOVA) performed with R (http://www.r-project.org).

For microarray data analysis, expression estimates were calculated from the raw cell files[17] obtained from ArrayExpress using GC robust multi-array average (gcRMA)[41]. Statistical testing for differential expression between the different data sets was performed with LIMMA[42]. Correction for multiple testing was done by computing Q-values[43]. These analyses were done under R using Bioconductor packages (http://www.bioconductor.org). By fixing the false discovery rate (FDR) to 1, we identified the following number of genes differentially expressed: FILp-CLV3p: 1379; FILp-WUSp: 1165; CLV3p-WUSp: 1357. Genes affected by protoplasting[17] (592 genes identified) were not considered in the analysis.

Electrophoresis mobility shift assays. The B3 DNA-binding domain of ARF5 (residues 120–274) was cloned into the vector pETM-11 (ref. 44). The protein was produced in *Escherichia coli* strain Rosetta 2 (Novagen) and purified on nickel sepharose high performance media (GE Healthcare) in a buffer containing 20 mM Tris-HCl pH 8.0 and 0.5M NaCl. For EMSA, single-stranded oligonucleotides were annealed to complementary oligonucleotides in annealing buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl and 1 mM EDTA). The dsDNA (4 pmol) with a protruding G was fluorescently labelled with Cy5-dCTP (8 pmol) (GE Healthcare) using 1 unit of Klenow fragment polymerase (Ozyme) in 1× Klenow buffer for 1 h at 37 °C. Binding reactions were performed with 1.5 μM of purified B3 domain and 25 nM of labelled dsDNA in 20 μl of binding buffer (20 mM HEPES-NaOH pH 7.9, 50 mM KCl, Tris-HCl 100 mM pH 8.0, 1% glycerol, 56 μg fish sperm DNA (Roche)). Binding reactions were loaded onto native 6% polyacrylamide gels and electrophoresis was conducted at 90 V for 75 min at 4 °C in 0.5× TBE (45 mM Tris, 45 mM boric acid and 1 mM EDTA pH 8.0). Gels were further scanned on a Typhoon 9400 scanner (excitation light 649 nm, emission filter 670 nm band-pass filter (670 BP 30); Molecular Dynamics).

The following oligonucleotides were used: TCA (control) 5′-ATACACGCAA**T GTCT**CCCTTT**TGTCTC**TTCCAC-3′; site A 5′-GCAAAGAAGCAT**GACA**TA CGAAT**GAGACA**ATTTCAGTTTT-3′; site A mutated 5′-GCAAAGAAGCAT **GCCA**TACGAAT**GAGCCA**ATTTCAGTTTT-3′; site B 5′-GTTATATGATTAT AACTT**GACA**GACCAAATAATCATCTTA-3′; site B mutated 5′-GTTATATG ATTATAACTT**GCCC**GCCCAAATAATCATCTTA-3′; site C 5′-AGCTGGTC T**GACA**GGGTACGCCGGT**TGTC**GGGAGGAAGAA-3′; site C mutated 5′-AG CTGGTCT**GCCA**GGGTACGCCGGT**TGGC**GGGAGGAAGAA-3′. The putative ARF binding sites are highlighted in bold and the mutations are underlined.

Cytology and cell biology. To analyse xylem defects in *ahp6-1* mutants expressing AHP6–GFP, AHP6–VENUS and AHP6-3×VENUS, roots of 5-day-old seedlings were cleared with chloral hydrate. RNA *in situ* hybridization on sections was performed as described[2] using full-length probes amplified by PCR. Whole-mount RNA *in situ* hybridization was performed as described[45] except that, before the treatment with proteinase K, permeability of the tissues was increased by digesting cell walls for 4 min at room temperature with an enzyme mix containing 0.5% Macerozyme R10 (Yakult Honsha), 0.5% cellulase RS (Yakult Honsha), 0.25%, pectolyase (Sigma), 0.75% pectinase (Serva) in water further diluted 6 times in PBS with 0.1% Tween20.

For PIN1 whole-mount immunolocalization, inflorescences were first fixed in FAA (5% formaldehyde, 50% ethanol and 10% acetic acid) for 1 h at room temperature (as all subsequent steps, unless specified) and dehydrated by a serial change of 70%, 90% and 100% ethanol (10 min each). Samples were then rehydrated by a serial change of 90%, 70%, 50%, 30% ethanol in microtubule-stabilising buffer (MTSB: 5 mM EGTA, 5 mM MgSO4, pH 7) plus 0.1% Triton and equilibrated with two washes of 10 min in MTSB. The enzyme mix used for whole-mount *in situ* RNA hybridization (see above) was diluted four times in MTSB and cell wall digestion was carried out for 45 min. Samples were then washed 3 times for 10 min in MTSB plus 0.1% Triton. Before antibody application, samples were pre-treated for 1 h in MTSB, 10% DMSO, 3% NP-40 then 1 h in solution A (MTSB, 0.1% Triton, 3% BSA from Sigma). Samples were then incubated overnight at 4 °C with the primary antibody (Ap20 anti-PIN1 antibody (Santa Cruz), dilution 1:100 in solution A), before being washed 4 times for 10 min in MTSB plus 0.1% Triton. Samples were then incubated with the secondary antibody (Alexa 488 antibody from Invitrogen, dilution 1:500 in solution A) for 3 h at 37 °C. The excess of antibody was then removed by 5 washes of 10 min in MTSB plus 0.1% Triton followed by 5 washes in water (10 min each). Samples were then observed using confocal microscopy.

Microscopy and live imaging. SEM was performed as previously described[2]. Cleared roots and RNA *in situ* hybridizations were observed with a transmission microscope under brightfield or differential interference contrast (DIC) illumination (Axio Imager 2, Zeiss). Confocal microscope observations were done either on a LSM-510 laser-scanning confocal microscope (Zeiss), a confocal spinning disc DMI400 microscope (Leica), a SP5 spectral detection confocal microscope (Leica) or on a LSM-710 spectral laser-scanning confocal microscope (Zeiss). Images were processed using ImageJ (http://rsbweb.nih.gov/ij/). Serial sections were used to count the number of cells expressing *pLFY::GFP* in the younger primordia. Culture and imaging of living SAMs expressing DR5::VENUS was performed as described previously[30] except for adding 555 nM *N*6-Benzyladenine (BAP; Duchefa) in the culture medium. The meristems were allowed to recover for 12 h before starting imaging every 12 h for 72 h. New primordia were scored as co-initiated when they exhibited similar number of VENUS-positive nuclei at a new time point.

Measurements and image analysis. All measurements and image analyses were done using ImageJ (http://rsbweb.nih.gov/ij/). Meristem width was measured on maximum-intensity projections of confocal serial images of living meristems after staining with FM4-64 (ref. 46). For each meristem, the youngest primordium P(n) separated from the meristem by a clear crease was selected. Then, the width was defined as the distance from this boundary to the opposite side of the meristem, located between P(n−1) and P(n−4) (See Supplementary Fig. 6a, b).

For measuring the size of the expression domain of GFP markers (*pWUS::GFP, DR5::VENUS*) in the meristem, images for a given marker were thresholded using the same fluorescence intensity range and transformed into binary images. The measure of the area of the central *DR5::GFP* negative region (see Supplementary Fig. 6i, j) was estimated by fitting the largest possible disc between existing primordia expressing DR5::GFP.

The minimal distance of organ initiation to the meristem centre and divergence angle at organ apparition was established using the coordinates of the meristem centre and coordinates of the centre of the initium as soon as the first DR5::VENUS-positive nuclei appeared during the kinetics. Organ displacement was deduced from the evolution of the organ-to-meristem centre distance during the time-lapse imaging. In this analysis, the centre of the meristem was defined as the minimal variation centre of the phyllotactic pattern as described[47]. To obtain the coordinates of the centre in a given image, we either used the FindCenter program (http://www.math.smith.edu/phyllo//Research/findcenter/findcenter.html) or we determined it manually by finding the optimal position giving divergence angles between all organs as close as possible to 137.5°. Both methods gave similar results, but we found that the manual determination was generally more robust. All distances were calculated on projections of serial confocal sections without taking into account the *z* coordinates.

The geometric determination of i1 and i2 positions on meristems was done first by determining the centre of the meristem as described above on a projection of serial confocal sections. Then the centres of all primordia from P4 till the oldest available (typically P8 to P14) were determined. The distance to the centre was then computed for all primordia in order to calculate the plastochron ratios (PR, which is the ratio of the distance to the centre of two consecutive primordia—older divided by younger). A mean PR was calculated for each plant. Then, positions of P3 to i2 were successively calculated at a 137.5° divergence angle from the previous organ (following the handedness of the spiral for the plant analysed) and using the mean PR of the plant to set the distance to the meristem centre. All calculations were performed using GeoGebra (http://www.geogebra.org/cms/fr/). Note that the fit of calculated positions of P3–P1 primordia with their actual positions provided an internal quality control for this geometrical modelling. In addition, the high predictive capacity of this method for positioning i1 was further demonstrated using 16 *DR5::VENUS* meristems (Supplementary Fig. 11).

For quantification of fluorescence intensities over a region (TCS::GFP or AHP6–GFP), we used summation-intensity projections in the region of interest and calculated the raw integrated density of the fluorescence. For TCS::GFP, we re-sliced the confocal stacks to obtain longitudinal slices of the region corresponding i2, i1 and P1. Fluorescence levels were then measured only from cells in L1 and L2. AHP6–GFP fluorescence levels at the predicted i1 and i2 sites (see above) were also calculated only from L1 and L2 cells over the area of a circle of 100 μm^2 centred on the predicted position of i1 and i2 (this corresponds approximately to the size of i1 when detected with *DR5::VENUS*). Profiles of AHP6 distribution were obtained using the 'Plot Profile' function of ImageJ along the regions indicated on Fig. 3h and Supplementary Fig. 12.

PIN1 transport network analysis. To map PIN1 in the meristems using immunolocalization images, cells were segmented using the Merryproj and Merrysim softwares[48] on the projection of a confocal stack. Influence zones were analysed as described[12]. STSE software was used to process segmentations and PIN1 orientations in order to obtain influence zones, and STSE and PlantGL were used to generate the colour maps[49,50].

To calculate the PIN1 polarity coherence index for a given cell, we computed all unit vectors pointing from the centre of this cell to the neighbouring cells with a side facing PIN1 proteins. We averaged all these vectors, and normalized the average, so as to obtain a unit vector defining the cell PIN1 polarity, which is representative of the direction in which auxin is transported in the cell of interest. To analyse the coherence of auxin transport, we defined for each cell the average of the polarity vectors of all neighbouring cells (including the cell of interest). The coherence index is the norm of this final average polarity vector. The coherence index has the value 1 when all cells have the same polarity, and a low value if polarities are very different. In order to generate a control, we replaced the measured PIN1 distribution by a distribution in which PIN1 proteins in a given cell are reallocated randomly to the other sides of each cell (and keeping the same number of sides carrying PIN1 so that connectivity remains the same). We recomputed the coherence index for this 'random' PIN1 distribution. Scenarios and index calculations were implemented using Python.

Measures of phyllotactic sequences. Measures of the sequences of silique divergence angles was performed as described[51]. For each phyllotactic measurement, several plants of the different genotype tested were grown in parallel (always including Col-0 and *ahp6-1* control individuals). For each individual of each genotype, the variance of the divergence angles was computed and individual variances of divergence angles were compared between genotype using a non-parametric Kruskal–Wallis test under R, as their distributions were not normally distributed.

Analyses of *ahp6-1 pAHP6::AHP6–VENUS* and *ahp6-1 pAHP6::AHP6–3×VENUS* were performed on T2 transformants which were hemizygote or homozygote for the transgene.

Models used for characterization of permutation patterns. To investigate the presence of particular motifs in phyllotactic sequences, we pooled four independent experiments of measurements with reproducible results, providing a data set of 82 wild-type and 89 *ahp6-1* plants. An exploratory analysis highlighted two characteristics of the divergence angle sequences: (1) the existence of short segments of non-canonical divergence angles along measured sequences (Fig. 1h, i, Supplementary Fig. 4c–f); and (2) almost all the possible angle values between 0 and 360° were observed with highest frequencies around the canonical angle α (Supplementary Fig. 4a). At least four classes of divergence angles were apparent but they were not unambiguously separated. To test if the segments of non-canonical angles could be explained by permutations and given the noisy character of the measurements, we designed a stepwise modelling approach[52,53] with two objectives: (1) to identify permutation patterns; and (2) optimally label the measured divergence angle sequences.

In a first step, a stationary hidden first-order Markov chain was estimated on the basis of the pooled measured divergence angle sequences (171 sequences representing a cumulative length of 5,220 angles). In this hidden first-order Markov chain, the states of the non-observable Markov chain represents 'theoretical' divergence angles whereas von Mises observation distributions attached to each state of the non-observable Markov chain represents measurement uncertainty. The von Mises distribution[54] is a univariate Gaussian-like periodic distribution for a variable $x \in [0,360^\circ)$. The von Mises observation distributions estimated for the five states of the non-observable Markov chain were centred on the multiples of the canonical divergence angle α, 2α, $-\alpha$, 3α, -2α (see Supplementary Table 2). The permutation of 2 consecutive organs generates the divergence angles 2α, $-\alpha$ and 3α. The identification of -2α using this five-state model suggested the occurrence of permutations involving 3 organs in the measured sequences. If in addition to 2-permutations, 3-permutations are considered, the divergence angles -2α, 4α and 5α are expected to be observed[52]. As the standard deviations of these von Mises observation distributions were quite similar, particularly for the most represented states corresponding to α, 2α and $-\alpha$, we chose to estimate a five-state hidden first-order Markov chain in which the von Mises observation distributions share the same concentration parameter (inverse variance). The optimally labelled divergence angle sequence (that is, discrete sequence with five possible values chosen among α, 2α, $-\alpha$, 3α, -2α) was then computed for each observed sequence using the estimated hidden first-order Markov chain.

In a second step, the memories of a variable-order Markov chain were optimally selected[55] on the basis of these labelled divergence angle sequences. This can be interpreted as a way to identify local dependencies between successive divergence angles. For the selection of these memories, we chose to discard the individuals that were very poorly explained by the estimated hidden first-order Markov chain (10 individuals out of 171 whose posterior probability of the optimally labelled divergence angle sequence <0.13). The variable-order Markov chain was a mixed first-/second-order Markov chain where the first-order memory 2α was replaced by the four second-order memories $\alpha 2\alpha$, $2\alpha 2\alpha$, $-\alpha 2\alpha$, $-2\alpha 2\alpha$ (the memory $3\alpha 2\alpha$ was not observed) with respect to a simple first-order Markov chain. This means that to predict accurately the most frequent permutation patterns, it is only necessary to take into account the divergence angle that precedes 2α. This is illustrated by the building of the 2-permutation pattern $[2\alpha\ -\alpha\ 2\alpha]$ as a succession of memories with high transition probabilities: $\alpha 2\alpha \overset{0.91}{\rightarrow} -\alpha \overset{0.72}{\rightarrow} -\alpha 2\alpha$ (Supplementary Table 3) instead of $2\alpha \overset{0.48}{\rightarrow} -\alpha \overset{0.71}{\rightarrow} 2\alpha$ with a simple first-order Markov chain. Finally, a hidden variable-order Markov chain was estimated where the underlying variable-order Markov chain has the memories previously selected.

One advantage of hidden Markov models is the capability to compute an absolute measure of the relevance of the optimally labelled divergence angle sequence as a posterior probability (that is, weight of this optimally labelled divergence angle sequence among all the possible labelled divergence angle sequences that can explain a given observed sequence). One shortcoming of hidden Markov models is that some multiples of the canonical divergence angle that occur rarely (for example, 4α, 5α) as well as alternative phyllotaxis (for example, Lucas with a canonical divergence angle of 99.5°) cannot be modelled. To be able to further investigate these sequences, we used a combinatorial mixture model as described[52]. The final results are then a consensus deduced from the divergence angle sequence optimally labelled by the hidden variable-order Markov chain and the combinatorial mixture model (Fig. 1k).

31. Deveaux, Y. *et al.* The ethanol switch: a tool for tissue-specific gene induction during plant development. *Plant J.* **36,** 918–930 (2003).
32. Yanai, O. *et al. Arabidopsis* KNOXI proteins activate cytokinin biosynthesis. *Curr. Biol.* **15,** 1566–1571 (2005).
33. Schlereth, A. *et al.* MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. *Nature* **464,** 913–916 (2010).
34. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743 (1998).
35. Lee, J.-Y. *et al.* Transcriptional and posttranscriptional regulation of transcription factor expression in *Arabidopsis* roots. *Proc. Natl Acad. Sci. USA* **103,** 6055–6060 (2006).
36. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **42,** 819–832 (2000).
37. Karimi, M., De Meyer, B. & Hilson, P. Modular cloning in plant cells. *Trends Plant Sci.* **10,** 103–105 (2005).
38. Udvardi, M. K., Czechowski, T. & Scheible, W.-R. Eleven golden rules of quantitative RT–PCR. *Plant Cell* **20,** 1736–1737 (2008).
39. Rieu, I. & Powers, S. J. Real-time quantitative RT–PCR: design, calculations, and statistics. *Plant Cell* **21,** 1031–1033 (2009).
40. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **26,** 509–515 (2004).
41. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99,** 909–917 (2004).

42. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*. 3, http://dx.doi.org/10.2202/1544-6115.1027 (2004).
43. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
44. Dümmler, A., Lawrence, A.-M. & de Marco, A. Simplified screening for the detection of soluble fusion constructs expressed in *E. coli* using a modular set of vectors. *Microb. Cell Fact.* **4,** 34 (2005).
45. Hejátko, J. *et al. In situ* hybridization technique for mRNA detection in whole mount *Arabidopsis* samples. *Nature Protocols* **1,** 1939–1946 (2006).
46. Grandjean, O. *et al. In vivo* analysis of cell division, cell growth, and differentiation at the shoot apical meristem in *Arabidopsis*. *Plant Cell* **16,** 74–87 (2004).
47. Hotton, S. Finding the center of a phyllotactic pattern. *J. Theor. Biol.* **225,** 15–32 (2003).
48. Barbier de Reuille, P. B., Bohn-Courseau, I., Godin, C. & Traas, J. A protocol to analyse cellular dynamics during plant development. *Plant J.* **44,** 1045–1053 (2005).
49. Stoma, S., Fröhlich, M., Gerber, S. & Klipp, E. STSE: Spatio-temporal simulation environment dedicated to biology. *BMC Bioinformatics* **12,** 126 (2011).
50. Pradal, C., Boudon, F., Nouguier, C., Chopard, J. & Godin, C. PlantGL: a Python-based geometric library for 3D plant modelling at different scales. *Graph. Models* **71,** 1–21 (2009).
51. Peaucelle, A., Morin, H., Traas, J. & Laufs, P. Plants expressing a *miR164*-resistant *CUC2* gene reveal the importance of post-meristematic maintenance of phyllotaxy in *Arabidopsis*. *Development* **134,** 1045–1050 (2007).
52. Refahi, Y. *et al.* in *Combinatorial Pattern Matching* (*Lecture Notes in Computer Science*) Vol. 6661 323–335 (Springer, 2011).
53. Guédon, Y. *et al.* Pattern identification and characterization reveal permutations of organs as a key genetically controlled property of post-meristematic phyllotaxis. *J. Theor. Biol.*, (2013).
54. Mardia, K. V. & Jupp, P. E. *Directional statistics* (Wiley, 2000).
55. Csiszár, I. & Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Theory* **52,** 1007–1016 (2006).

LETTER

doi:10.1038/nature12679

Chasing acyl carrier protein through a catalytic cycle of lipid A production

Ali Masoudi[1], Christian R. H. Raetz[1]‡, Pei Zhou[1] & Charles W. Pemble IV[2,3]

Acyl carrier protein represents one of the most highly conserved proteins across all domains of life and is nature's way of transporting hydrocarbon chains *in vivo*. Notably, type II acyl carrier proteins serve as a crucial interaction hub in primary cellular metabolism[1] by communicating transiently between partner enzymes of the numerous biosynthetic pathways[2,3]. However, the highly transient nature of such interactions and the inherent conformational mobility of acyl carrier protein[2] have stymied previous attempts to visualize structurally acyl carrier protein tied to an overall catalytic cycle. This is essential to understanding a fundamental aspect of cellular metabolism leading to compounds that are not only useful to the cell, but also of therapeutic value. For example, acyl carrier protein is central to the biosynthesis of the lipid A (endotoxin) component of lipopolysaccharides in Gram-negative microorganisms, which is required for their growth and survival[4,5], and is an activator of the mammalian host's immune system[6,7], thus emerging as an important therapeutic target[8–10]. During lipid A synthesis (Raetz pathway), acyl carrier protein shuttles acyl intermediates linked to its prosthetic 4′-phosphopantetheine group[2] among four acyltransferases, including LpxD[11]. Here we report the crystal structures of three forms of *Escherichia coli* acyl carrier protein engaging LpxD, which represent stalled substrate and liberated products along the reaction coordinate. The structures show the intricate interactions at the interface that optimally position acyl carrier protein for acyl delivery and that directly involve the pantetheinyl group. Conformational differences among the stalled acyl carrier proteins provide the molecular basis for the association–dissociation process. An unanticipated conformational shift of 4′-phosphopantetheine groups within the LpxD catalytic chamber shows an unprecedented role of acyl carrier protein in product release.

Although the paradigm for acyl carrier protein (ACP) association with protein partners is thought to be exceedingly transient, the LpxD acyltransferase in the Raetz pathway (Supplementary Fig. 1a) binds ACP with very high affinity ($K_d = 59$ nM)[12]. LpxD transfers *R*-3-hydroxymyristoyl (β-OH-C14) acyl chains that are delivered by ACP to the amino group of uridine diphosphate 3-O-(β-OH-C14)-α-D-glucosamine (UDP-acyl-GlcN)[12] (Supplementary Fig. 1b). Moreover, LpxD follows an ordered sequential kinetic mechanism in which acyl-ACP binds first and, importantly, holo-ACP (acyl chain has been liberated) dissociates last[12]. Given the canonical mode of ACP-mediated transfer often allows the rapid exchange between protein partners[13], such pronounced association between ACP and LpxD is most unusual and instead suggests a 'strong, transient'[14] interaction that would require a yet to be identified 'molecular trigger' for dissociation. Thus, we reasoned that the crystallographic study of the ACP–LpxD complex offers a unique opportunity to gain the detailed molecular basis for ACP-based acyl delivery and a deeper understanding of protein–protein communication more generally.

We present here three X-ray co-crystal structures of ACP bound to LpxD, each of which captured a different form of the carrier protein: intact-acyl-ACP, hydrolysed-acyl-ACP and holo-ACP (Fig. 1a, b, Supplementary Fig. 2 and Supplementary Table 1). These structures were resolved to 2.1, 2.9 and 2.15 Å resolution, respectively (Supplementary Table 2). In each case, the triclinic unit cell contained two LpxD trimers with differing bound states of ACP (Supplementary Fig. 3). The overall B values for all ACPs modelled in each structure are two- to threefold higher than LpxD (Supplementary Fig. 4 and Supplementary Table 2) and, accordingly, the observed electron density was weaker in some regions (Supplementary Fig. 3). Nonetheless, the placement of side-chains at the protein–protein interface, prosthetic groups, and acyl chain(s) were apparent (Supplementary Fig. 2d–f and Supplementary Fig. 5). Our structures show the 4′-phosphopantetheine group (4′-PPT) (attached to Ser 36) and its β-OH-C14 acyl chains having vacated the canonical hydrophobic cavity extending through the core of ACP[2,15–17]—all of which require considerable movement (Supplementary Fig. 6). The overall architecture of the LpxD trimer is similar to previously reported X-ray structures[18,19] in that each monomer of LpxD can be subdivided into three domains (Fig. 1c and Supplementary Fig. 2): the amino (N)-terminal uridine-binding domain, which is tethered to the left-handed β-helix domain that harbours the conserved catalytic His 239 residue[12], and a carboxy (C)-terminal domain.

The structures show three molecules of the carrier protein are localized to the C-terminal end of LpxD (Fig. 1b). Notably, we have identified the ACP recognition domain (ARD) (Fig. 1c), which is formed by the C-terminal domain and the last beta-coil of the left-handed β-helix domain, providing the molecular basis for ACP association. This contrasts with a previous study that suggested the uridine-binding domain as the likely ACP docking site owing to its proximity to the catalytic cleft[18]. Although the analogous C-terminal region of the LpxA acyltransferase[20] is found to adopt a completely different orientation from that of LpxD, it may serve a similar function in binding ACP (Supplementary Fig. 7).

By virtue of the complete engagement of ACP, three competent active sites are created (Fig. 2a). Each ACP–LpxD interface buries a surface area of approximately 530 Å^2 and is predominated by complementary electrostatic interactions (Fig. 2b). In addition, van der Waals contacts and extensive interaction with the prosthetic group contribute to the large binding footprint that explains the 'strong, transient' nature of these two protein partners. A combination of residues located on the 'universal recognition helix' (helix II)[21], as well as portions of L1, L2 and helix-III of ACP, provide the acidic surface that binds a pronounced basic patch on LpxD. This surface feature of ACP can be subdivided into two highly acidic regions, I and II, which include residues Glu 30–Met 44 and Ala 45–Glu 60, respectively. The complementary binding surface on LpxD involves residues from all three monomers (denoted by prime symbols) and forms a shallow groove between coiled coils of the ARD into which helix-II packs (Supplementary Fig. 8).

Within region I, Asp 35, Ser 36, Leu 37, Asp 38, Val 40, Glu 41 and Met 44 are important for binding the N-terminal end of the recognition

[1]Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710, USA. [2]Human Vaccine Institute, Duke University Medical Center, Durham, North Carolina 27710, USA. [3]Duke Macromolecular Crystallography Center, Duke University Medical Center, Durham, North Carolina 27710, USA.
‡Deceased.

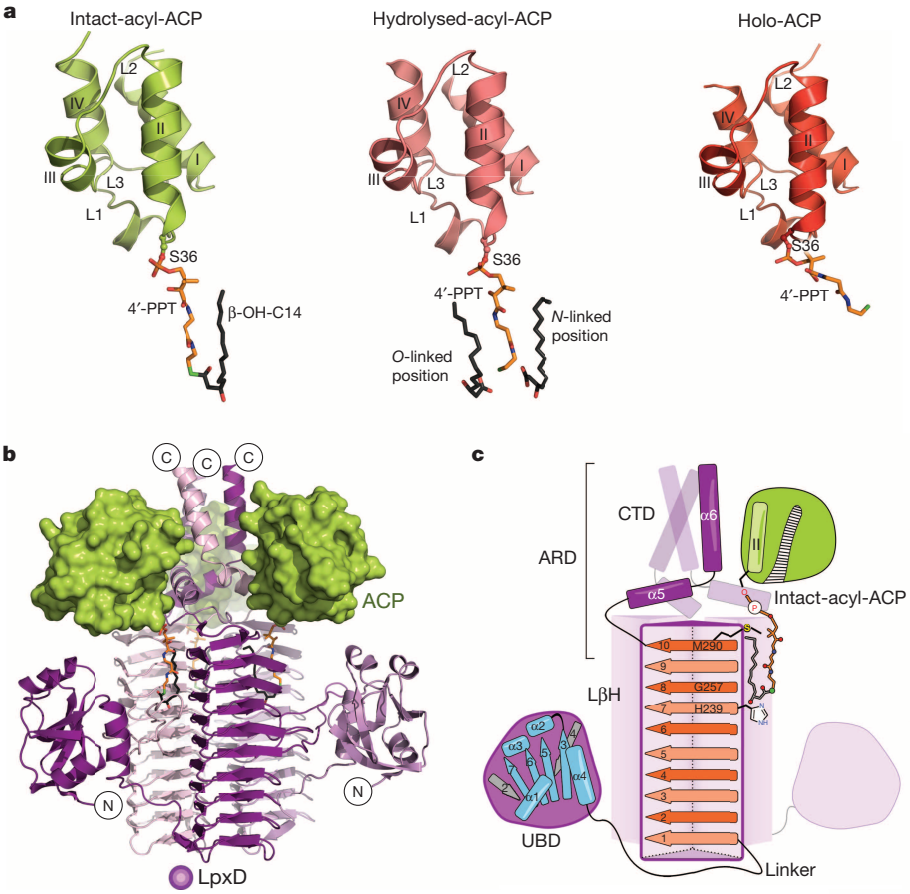


Figure 1 | Stalled ACPs bound to LpxD. **a**, The three forms of ACP: intact-acyl-ACP (green), hydrolysed-acyl-ACP (salmon) and holo-ACP (red). ACP adopts a four-helix bundle (I–IV) with several loop regions (L1–L3)[28,29]. The 4′-PPT arm and acyl chains are shown as stick models and coloured orange or dark grey, respectively, and by atom. **b**, Overall architecture of the ACP–LpxD complex (intact-acyl-ACP shown). Three ACPs bind the C-terminal end of the LpxD trimer (coloured by chain purple/magenta/light pink). **c**, Cartoon rendering showing the overall fold and interaction of acyl-ACP with LpxD. Highlighted dark purple is a single monomer of LpxD with subdomains indicated: uridine-binding domain (UBD), left-handed β-helix domain (LβH) and C-terminal domain (CTD). ACP, interfacing with the ARD, and its acyl-4′-PPT group are shown. The locations of the catalytic base (His 239), oxyanion hole (Gly 257), and molecular 'hydrocarbon ruler' (Met 290) are indicated within the ten β-helical coils (orange strands) of a single left-handed β-helix domain.

helix to the base of the ARD domain (Supplementary Fig. 8), and the interactions were notably present in all three structures. Region II of ACP interacts with the upper portion of the ARD domain, the details of which differ substantially among the three stalled ACP complexes (discussed below). Most of the residues within regions I and II are conserved among other type II carrier proteins (Supplementary Fig. 2b) and have been implicated as key modulators of ACP association[22,23]. The most universal electrostatic interaction shown across the intact-, hydrolysed- and holo-ACP structures is between Arg 293 of LpxD and Asp 35, Asp 38 and Glu 41 of ACP that flank Ser 36. This coordination point is crucial for ACP association with LpxD as mutation of Arg 293 to alanine results in a 23-fold increase in Michaelis constant (K_m) for acyl-ACP compared with wild-type LpxD, with little effect on catalyst rate constant (k_{cat})[12]. These interactions orient ACP by positioning the pantetheinylated Ser 36 residue towards the catalytic chamber of LpxD.

In addition to the protein–protein interactions, the prosthetic group and acyl chain of intact-acyl-ACP extensively contact the surface of the partner enzyme (Supplementary Fig. 9). Inspection of the electron density indicates that the thioester scissile bond remained unbroken, which required a His239Ala LpxD variant (Supplementary Fig. 5a). The β-OH-C14 acyl chain and the 4′-PPT arm adopt a horseshoe-like conformation, which in effect buries the acyl chain between the prosthetic group and a pronounced hydrophobic channel (*N*-channel) formed between LpxD monomers (Fig. 2c). The 4′-phosphate moiety of 4′-PPT is directly coordinated by Asn 310 and Arg 314 of LpxD and may partly contribute to some level of specificity. The remainder of the 4′-PPT arm adopts a rather extended conformation, stretching over 14 Å, and interacts with residues primarily along the rim of the *N*-channel. This conformation places the thioester bond in proximity to the alanine-substituted His 239 catalytic base and orients the carbonyl-oxygen of β-OH-C14 towards the amide nitrogen atom of Gly 257′′, corroborating its role in forming the oxyanion hole[12,24].

Two features of LpxD specificity towards β-hydroxyacyl chains are explained by the intact-acyl-ACP structure. First, the terminal two carbon atoms of β-OH-C14 pack against Met 290′′ located at the far end of the *N*-channel (Fig. 2c and Supplementary Fig. 9). Given that LpxD is highly specific for 14-carbon acyl-chain-lengths[12], this supports the role of Met 290 as a 'hydrocarbon ruler'[19]. Second, LpxD is particularly selective for the β-hydroxyl group of the acyl chain; its removal completely abrogates acyltransferase activity[25]. Our structures show an intricate hydrogen bond network between the β-hydroxyl group and Asp 232′, Gln 236′ and a critical water-mediated bridge with Asp 216′′ and the main-chain nitrogen atoms of Phe 183′′ and His239Ala′′.

By contrast to the intact-acyl-ACP complex, the thioester bond is broken in the hydrolysed-acyl-ACP structure (Fig. 2d, Supplementary Fig. 5b and Supplementary Fig. 9). Notably, we discovered a second molecule of β-OH-C14 fatty acid bound to the LpxD surface (Supplementary Fig. 10), which shows an extra hydrophobic channel (*O*-channel) that probably binds the acyl chain that is ester-linked to the UDP-acyl-GlcN lipid substrate (Supplementary Fig. 1b). A superposition with the previously reported *Chlamydia trachomatis* LpxD in complex with UDP-GlcNAc[18] illustrates the proximity of the carboxylate head group of the extra fatty acid to the anticipated binding locale of the 3-hydroxyl position of the GlcN ring (Supplementary Fig. 11).

What is striking about both the intact- and hydrolysed-acyl-ACP structures is the conformational similarity between the 4′-PPT prosthetic groups that are fully extended (Fig. 2c, d), whereas the holo-ACP structure shows an alternative conformation of its prosthetic group packing against the far end of the *N*-channel (Fig. 2e and Supplementary Fig. 9). Notably, in both intact- and hydrolysed-acyl-ACP the pantetheinyl arms completely enclose the reaction chambers (Fig. 3a). Although this architecture probably stabilizes substrate binding within the hydrophobic *N*-channel, it raises a key question. Given a sequential

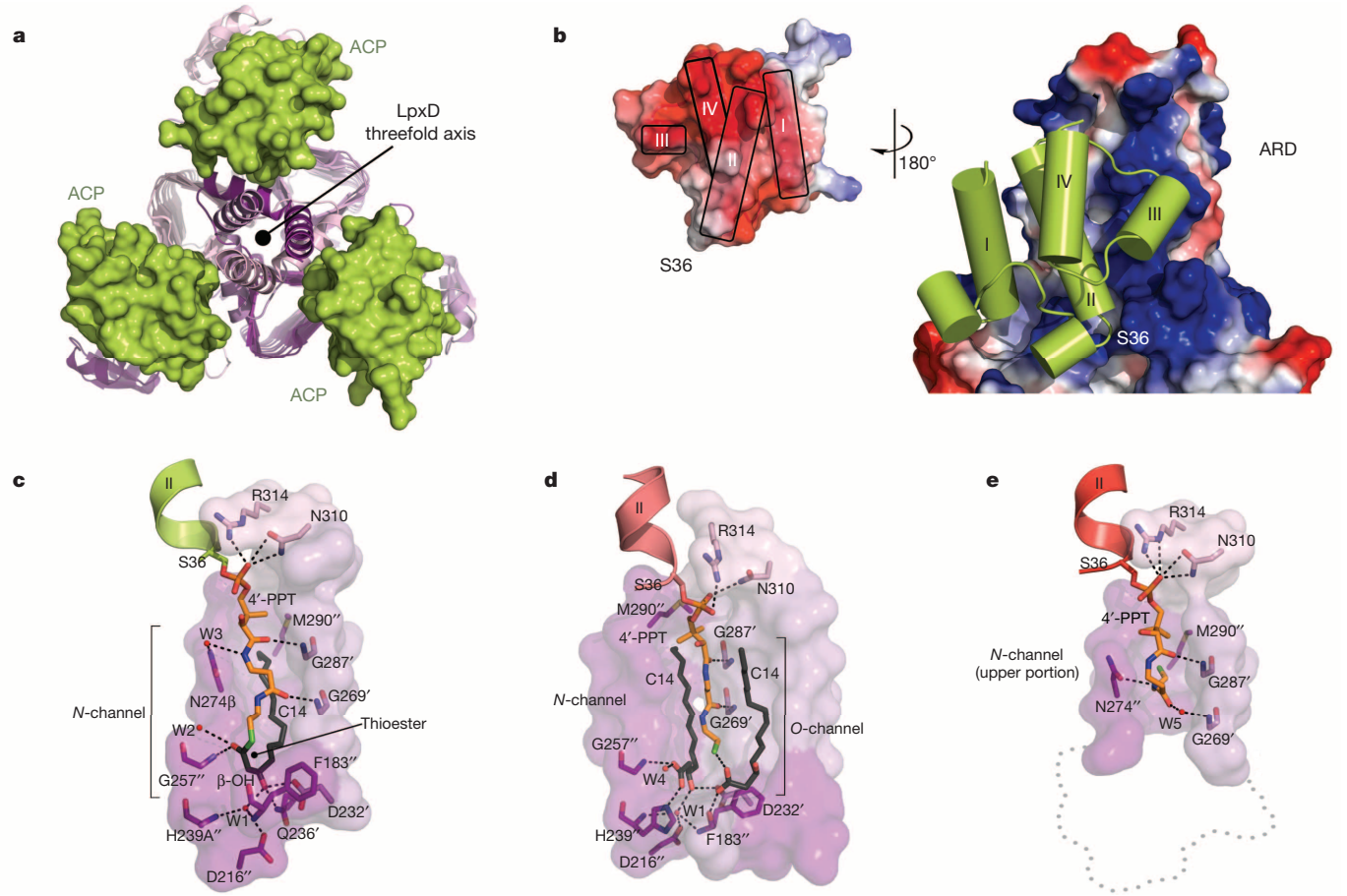


Figure 2 | Intermolecular interactions between ACP and LpxD. **a**, **b**, Overview of the protein–protein interactions (intact-acyl-ACP depicted). **a**, Top-down view of the ACP–LpxD complex showing three molecules of ACP bind per LpxD trimer. **b**, Electrostatic surface representation of the ARD and ACP (inset); the potential contours were scaled to +79.2 (blue) and −79.2 (red) k_BTe^{-1} (where k_B is the Boltzmann constant, T is temperature and e^{-1} is the charge of an electron). **c**–**e**, Detailed interactions between the LpxD reaction chamber and the bound acyl/4′-PPT groups of ACP. Hydrogen bonds are shown as black dashes. Molecular surfaces are of only those residues that contribute to interactions. **c**, Intact-acyl-ACP complex. The β-OH-C14 acyl chain delivered by ACP binds the hydrophobic *N*-channel and its terminal carbon atoms pack near Met 290′′. **d**, Hydrolysed-acyl-ACP complex. An equivalent β-OH-C14 acyl chain is shown bound to a newly identified hydrophobic channel (*O*-channel). **e**, Holo-ACP complex. The 4′-PPT arm interacts at the far end of the *N*-channel, positioning the terminal thiol near Met 290′′. A 'ghost outline' of the catalytic cleft is indicated by a grey dotted line.

ordered reaction mechanism, how does the di-acylated-GlcN product leave LpxD before holo-ACP if the reaction chamber is completely blocked? A structural comparison between intact- and holo-ACP sheds light on this matter by exposing a substantial movement that the 4′-PPT arm has undergone (Fig. 3b). Notably, the terminal thiol of the 4′-PPT has vacated the catalytic cleft and moves approximately 15 Å to be situated near Met 290. A sizeable region of the reaction chamber closest to the catalytic cleft is now open to solvent, thereby giving the diacyl-GlcN product an opportunity to dissociate before the release of holo-ACP.

The remarkable difference in conformation shown by the 4′-PPT arm, together with the ordered sequential mechanism of acyl transfer, prompted us to propose that this 'swing' motion may in fact be involved in 'triggering' the release of lipid product. Because every terminal thiol of holo-ACP was positioned approximately 3.7 Å away from Met 290 of LpxD, we considered the possibility of mutating this residue to a cysteine in an effort to induce a mixed-disulphide linkage post-catalysis. Biochemically, the Met 290Cys mutation abrogates acyl transfer to UDP-acyl-GlcN compared with wild-type LpxD (Supplementary Fig. 12). This suggests that a covalent bond is formed between the cysteinyl and 4′-PPT thiols. Thus, we reasoned that the addition of reducing agent would rescue acyl transfer. Accordingly, titration of dithiothreitol (DTT) into the reaction mixture recovers activity of Met290Cys-LpxD to levels indistinguishable from that of wild-type enzyme (Supplementary Fig. 12). These data suggest that the observed 4′-PPT motion has functional relevance in the course of product release. Moreover, as typically exemplified with 'strong, transient' interactions[14], this substantial movement of the pantetheine arm probably serves as the 'molecular trigger' that promotes the collapse of the ACP–partner complex.

Because our structures show different states of ACP stalled at the LpxD active site, an alignment of LpxD domains allows us to visualize the movements within ACP as it relates to the overall catalytic cycle (Fig. 3c), which to the best of our knowledge is unprecedented. Helix-I, helix-II and portions of L1 remain relatively unchanged (root mean squared deviation of approximately 1 Å) and are preserved for the purpose of remaining docked. The largest differences occur downstream of the recognition helix (Fig. 3c), including helices III, IV, L2 and L3 (root mean squared deviation of approximately 3 Å) and are an indication of what interactions must break for dissociation. A closer inspection shows that the intact-complex makes electrostatic interactions with the entire ARD interface that involves region II, whereas both the hydrolysed- and holo-ACP structures do not (Fig. 3d). These data indicate that the more extensive interactions are important for molecular recognition during association; however, after acyl transfer it is conformational changes in ACP that ultimately destabilize the protein–protein complex. In this context, one particular molecule of

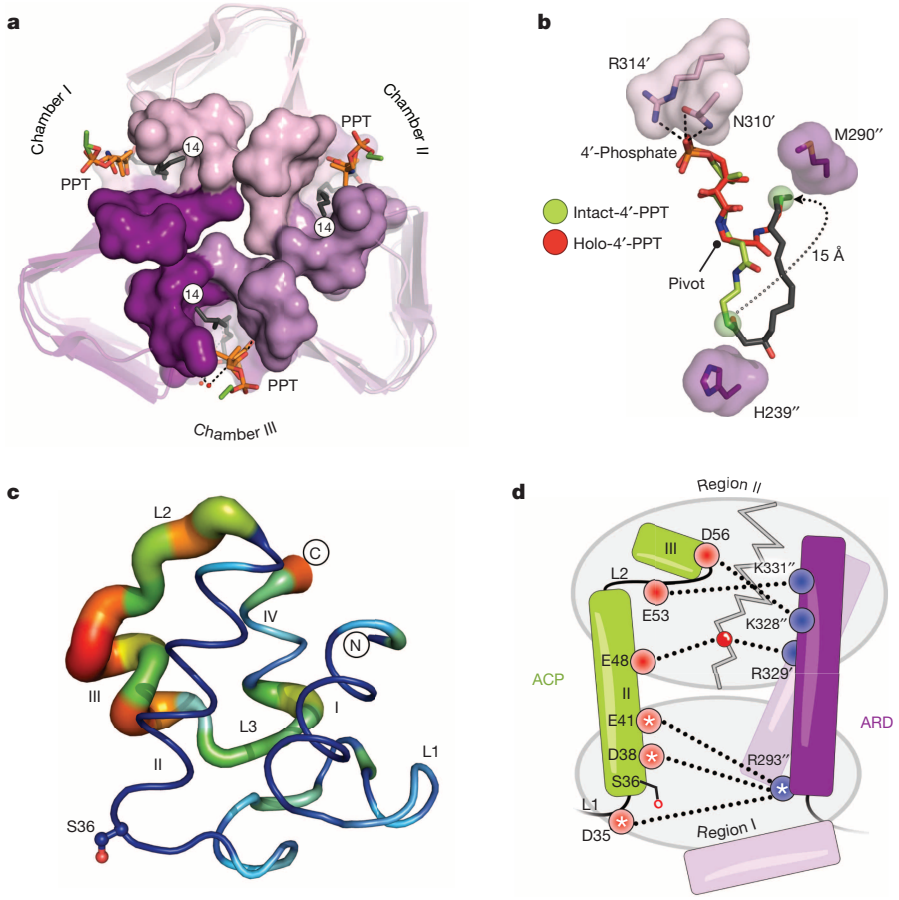


Figure 3 | ACP conformations and reorganization of its prosthetic group. **a**, Top-down view of LpxD (intact-acyl-ACP depicted) showing three reaction chambers enclosed by the 4′-PPT. **b**, Structural comparison between the intact- (green) and holo-4′-PPT (red) prosthetic groups. The 4′-PPT rearrangement is indicated (grey dotted arrow). **c**, Difference distance matrix calculated between fully modelled intact- (chain K) and holo-ACP (chain G). Deviations between like atoms are shown as a putty-sausage representation. Both the thickness and heat-map colouring indicate regions of least (thin, blue) to highest (thick, red) displacement. **d**, Schematic summarizing differences in electrostatic interactions between ACP complexes. Residues are indicated as red (acidic) or blue (basic) circles and subdivided according to region I or II affiliation. Interactions made by intact-acyl-ACP alone are depicted (dotted lines) and those common among all complexes are represented with a star. The grey zigzag indicates those electrostatic interactions that are broken in the hydrolysed- and holo-ACP product complexes.

holo-ACP within the asymmetric unit is displaced from the others and possibly represents a different binding state showing extra interactions that must break for dissociation (Supplementary Fig. 13).

The structures captured in this study begin to establish key molecular movements within ACP that initiate molecular recognition and a mechanism with which associations can be ultimately broken. In this context, we present a model for ACP-based synthesis of lipid A precursors catalysed by LpxD (Fig. 4) and can begin to extract some general principles for these types of interaction. First, although there does not yet seem to be a consensus-sequence binding motif on the surface

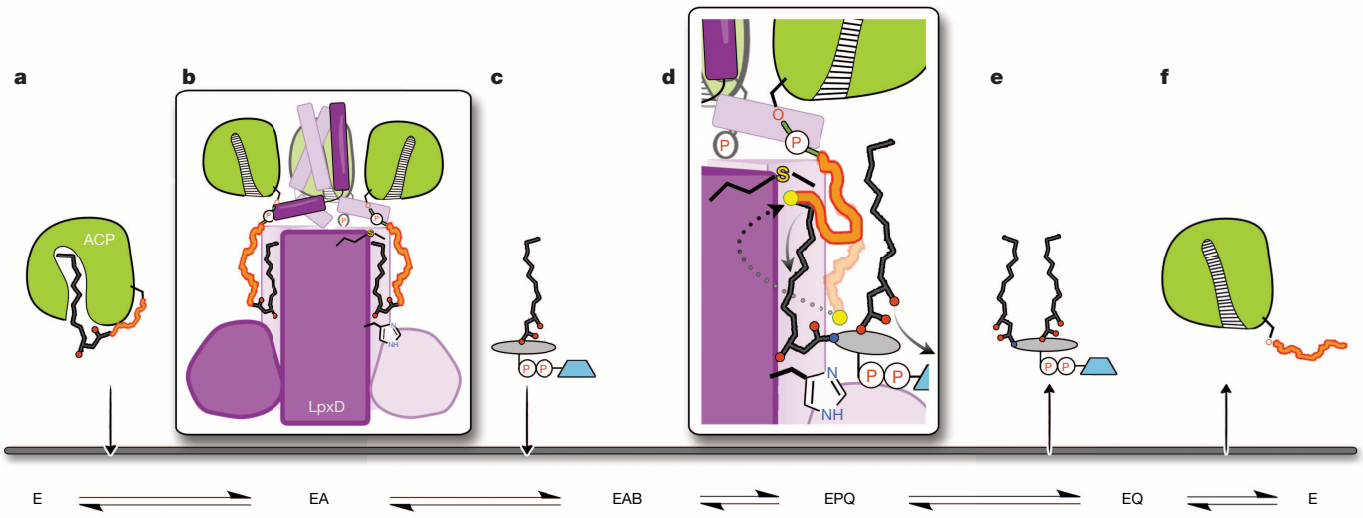


Figure 4 | Molecular basis for the ordered-sequential reaction mechanism and involvement of ACP in lipid-product release. **a**, **b**, Acyl-ACP (**a**) binds first to free LpxD forming the binary complex (**b**). ACP associates with the ARD and the acyl-4′-PPT packs into the hydrophobic *N*-channel. **c**, UDP-acyl-GlcN binds next, which initiates acyl transfer. **d**, In the ternary product complex the 4′-PPT arm of hydrolysed-acyl-ACP (transparent orange squiggle) completely encloses the reaction chamber, blocking UDP-diacyl-GlcN from leaving. By moving the 4′-PPT (dark orange squiggle) towards Met 290 (dotted arrow), the catalytic chamber opens up. **e**, **f**, This motion drives the eventual release of UDP-diacyl-GlcN (**e**) and 'triggers' conformational changes downstream of helix-II leading to holo-ACP dissociation (**f**).

of partner proteins, both regions I and II of ACP seem to be frequently involved, with most of the small number of other ACP-based complexes determined showing that helices II, -III, L1 and L2 are consistently used[22,23,26,27]. Second, because these regions are broadly universal, exploiting specific residue contacts at the interface and fine-tuning the buried surface area at these regions probably dictate how ACP-based associations can be both specific and temporary. For instance, our structures show the first direct, electrostatic coordination with the 4′-phosphate of both acyl- and holo-ACP, which in addition to extensive protein–protein contacts provide the 'strong, transient' nature of the ACP–LpxD complex. Third, because we observe ACP tied to an entire catalytic cycle, a substantial conformational change downstream of the recognition helix-II may represent a more general communication mechanism for breaking ACP–partner complexes. Finally, the stalled ACPs captured herein have shown an unprecedented role for carrier proteins in product release, and the contribution of the pantetheinyl group in both the formation and dissociation of the ACP–partner complex. Perhaps motion of the pantetheinyl group in other ACP-based complexes provokes dissociation in a similar manner, especially when considering other 'strong, transient' protein partners.

METHODS SUMMARY

Key to capturing ACP bound to LpxD at different stages of catalysis was the ability to preload *E. coli* ACP with its 4′-PPT prosthetic group by co-expression with holo-ACP synthase (ACPS), and subsequently charging holo-ACP with *R*-3-hydroxymyristic acid enzymatically using the soluble form of acyl-ACP synthetase from *Vibrio harveyi*. Holo-ACP was purified by nickel affinity, anion exchange, covalent chromatography, which exploits disulphide bond formation between the 4′-PPT group and ThioPropyl Sepharose resin (Sigma-Aldrich), and finally size-exclusion chromatography. Complete conversion of holo-ACP to its acylated form was confirmed by mass spectrometry and 2.5 M urea gel electrophoresis. The resulting acyl-ACP mixture was purified to homogeneity using size-exclusion chromatography. *E. coli* LpxD was overexpressed, and purified as previously reported[12]. An active site mutant of LpxD was also necessary to preserve the thioester bond of acyl-ACP and was generated by introducing an alanine substitution for His 239. Crystals of the complexes were grown by vapour diffusion and incubating either holo- or acyl-ACP with LpxD in a 1:1 (mol:mol) stoichiometry. Diffraction data were collected on the SER-CAT 22-BM and 22-ID beamlines at the Advanced Photon Source at Argonne National Laboratory. The structures were solved by the molecular replacement method using the published *E. coli* LpxD structure (Protein Data Bank 3EH0)[19] as the search model. *E. coli* ACP was intentionally omitted during molecular replacement and, instead, was manually rebuilt into unbiased difference electron density maps. The rate of LpxD catalysed conversion of [α-^{32}P]UDP-3-acylGlcN to [α-^{32}P]UDP-2,3-diacylGlcN was measured as described previously using thin-layer chromatography[25]. Details of all procedures are presented in Supplementary Information.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 April; accepted 17 September 2013.
Published online 6 November 2013.

1. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433,** 531–537 (2005).
2. Byers, D. M. & Gong, H. Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. *Biochem. Cell Biol.* **85,** 649–662 (2007).
3. White, S. W., Zheng, J., Zhang, Y. M. & Rock. The structural biology of type II fatty acid biosynthesis. *Annu. Rev. Biochem.* **74,** 791–831 (2005).
4. Galloway, S. M. & Raetz, C. R. H. A mutant of *Escherichia coli* defective in the first step of endotoxin biosynthesis. *J. Biol. Chem.* **265,** 6394–6402 (1990).
5. Belunis, C. J., Clementz, T., Carty, S. M. & Raetz, C. R. H. Inhibition of lipopolysaccharide biosynthesis and cell growth following inactivation of the kdtA gene in *Escherichia coli*. *J. Biol. Chem.* **270,** 27646–27652 (1995).
6. Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science* **282,** 2085–2088 (1998).
7. Akira, S., Uematsu, S. & Takeuchi, O. Pathogen recognition and innate immunity. *Cell* **124,** 783–801 (2006).
8. Onishi, H. R. *et al.* Antibacterial agents that inhibit lipid A biosynthesis. *Science* **274,** 980–982 (1996).
9. Williams, A. H., Immormino, R. M., Gewirth, D. T. & Raetz, C. R. H. Structure of UDP-N-acetylglucosamine acyltransferase with a bound antibacterial pentadecapeptide. *Proc. Natl Acad. Sci. USA* **103,** 10877–10882 (2006).
10. Jenkins, R. J. & Dotson, G. D. Dual targeting antibacterial peptide inhibitor of early lipid A biosynthesis. *ACS Chem. Biol.* **7,** 1170–1177 (2012).
11. Raetz, C. R. H., Reynolds, C. M., Trent, M. S. & Bishop, R. E. Lipid A modification systems in Gram-negative bacteria. *Annu. Rev. Biochem.* **76,** 295–329 (2007).
12. Bartling, C. M. & Raetz, C. R. H. Steady-state kinetics and mechanism of LpxD, the *N*-acyltransferase of lipid A biosynthesis. *Biochemistry* **47,** 5290–5302 (2008).
13. Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316,** 288–290 (2007).
14. Nooren, I. M. & Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **22,** 3486–3492 (2003).
15. Chan, D. I. & Vogel, H. J. Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430,** 1–19 (2010).
16. Ploskon, E. *et al.* Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. *Chem. Biol.* **17,** 776–785 (2010).
17. Roujeinikova, A. *et al.* Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J. Mol. Biol.* **365,** 135–145 (2007).
18. Buetow, L., Smith, T. K., Dawson, A., Fyffe, S. & Hunter, W. N. Structure and reactivity of LpxD, the *N*-acyltransferase of lipid A biosynthesis. *Proc. Natl Acad. Sci. USA* **104,** 4321–4326 (2007).
19. Bartling, C. M. & Raetz, C. R. H. Crystal structure and acyl chain selectivity of *Escherichia coli* LpxD, the *N*-acyltransferase of lipid A biosynthesis. *Biochemistry* **48,** 8672–8683 (2009).
20. Raetz, C. R. H. & Roderick, S. L. A left-handed parallel β helix in the structure of UDP-*N*-acetylglucosamine acyltransferase. *Science* **270,** 997–1000 (1995).
21. Frederick, A. F., Kay, L. E. & Prestegard, J. H. Location of divalent ion sites in acyl carrier protein using relaxation perturbed 2D NMR. *FEBS Lett.* **238,** 43–48 (1988).
22. Cryle, M. J. & Schlichting, I. Structural insights from a P450 carrier protein complex reveal how specificity is achieved in the P450(Biol) ACP complex. *Proc. Natl Acad. Sci. USA* **105,** 15696–15701 (2008).
23. Parris, K. D. *et al.* Crystal structures of substrate binding to *Bacillus subtilis* holo-(acyl carrier protein) synthase reveal a novel trimeric arrangement of molecules resulting in three active sites. *Structure* **8,** 883–895 (2000).
24. Kraut, D. A., Carroll, K. S. & Herschlag, D. Challenges in enzyme mechanism and energetics. *Annu. Rev. Biochem.* **72,** 517–571 (2003).
25. Kelly, T. M., Stachula, S. A., Raetz, C. R. & Anderson, M. S. The firA gene of *Escherichia coli* encodes UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine N-acyltransferase. The third step of endotoxin biosynthesis. *J. Biol. Chem.* **268,** 19866–19874 (1993).
26. Agarwal, V., Lin, S., Lukk, T., Nair, S. K. & Cronan, J. E. Structure of the enzyme-acyl carrier protein (ACP) substrate gatekeeper complex required for biotin synthesis. *Proc. Natl Acad. Sci. USA* **109,** 17406–17411 (2012).
27. Babu, M. *et al.* Structure of a SLC26 anion transporter STAS domain in complex with acyl carrier protein: implications for *E. coli* YchM in fatty acid metabolism. *Structure* **18,** 1450–1462 (2010).
28. Holak, T. A., Nilges, M., Prestegard, J. H., Gronenborn, A. M. & Clore, G. M. Three-dimensional structure of acyl carrier protein in solution determined by nuclear magnetic resonance and the combined use of dynamical simulated annealing and distance geometry. *Eur. J. Biochem.* **175,** 9–15 (1988).
29. Roujeinikova, A. *et al.* X-ray crystallographic studies on butyryl-ACP reveal flexibility of the structure around a putative acyl chain binding site. *Structure* **10,** 825–835 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge our co-author Christian R. H. Raetz, who shaped the lipid field with his curiosity and efforts, introducing many scientists to the field during his renowned career. We thank R. Brennan and W. Todd Lowther for reviewing the manuscript. Finally, we thank Z. Guan for the help with the mass spectrometry of ACP, H.-S. Chung and other members of Raetz laboratory, as well as J. M. Burg, for discussions. Crystal screening, data collection and data processing were conducted in collaboration with the Duke Macromolecular X-ray Crystallography Shared Resource. Diffraction data were collected remotely at the Southeast Regional Collaborative Access Team 22-BM and 22-ID beamlines at the Advanced Photon Source, Argonne National Laboratory, supported by the US Department of Energy, Office of Science and the Office of Basic Energy Sciences under Contract number W-31-109-Eng-38. This work was supported by National Institutes of Health grants GM-51310 and AI-055588 awarded to C.R.H.R. and P.Z.

Author Contributions A.M., C.R.H.R. and C.W.P. designed research; A.M. performed all biochemical experiments under the guidance of C.R.H.R., P.Z. and C.W.P.; A.M. performed all protein expression, purification and crystallization; A.M. and C.W.P. contributed to data collection, structure solution and refinement; A.M., C.R.H.R. and C.W.P. analysed and interpreted the structures; A.M. and C.W.P. made the figures and wrote the manuscript; A.M., C.R.H.R., P.Z. and C.W.P. discussed the results and commented on the manuscript.

Author Information Molecular coordinates and structure factors of intact-acyl-ACP, hydrolysed-acyl-ACP and holo-ACP in complex with LpxD have been deposited in the Protein Data Bank under accession numbers 4IHF, 4IHG and 4IHH, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.W.P. (charles.pemble@duke.edu).

METHODS

Expression and purification of His_6-LpxD. *E. coli* LpxD was overexpressed, and purified as previously reported[12]. Briefly, His_6-LpxD was expressed in *E. coli* Rosetta/pLysS. The membrane-free fraction was loaded onto a 5 ml Ni-NTA (Qiagen) column and eluted in one step with 200 mM imidazole. The His_6-tag was left intact and the resulting LpxD was concentrated to approximately 4–10 ml and loaded onto a High Load 26/60 Superdex 200 gel filtration column (GE) equilibrated with 10 mM Tris-HCl, pH 7.5, 500 mM NaCl and 1 mM DTT. The protein peak had an elution profile consistent with that of the LpxD homotrimer. Fractions were pooled and concentrated to approximately 40 mg ml^{-1} and stored in aliquots at −80 °C. The wild-type LpxD enzyme used in the LpxD assay was purified and stored in the absence of DTT.

Construction, expression and purification of His_6-LpxD point mutants. Site-directed mutants designed to alter Met 290 (Met290Cys), and His 239 (His239Ala) were accomplished using the QuikChange PCR mutagenesis kit (Stratagene). The LpxD variants were recombinantly expressed in C41(DE3) *E. coli* cells and purified in the same manner as described for wild-type His_6-LpxD. The His239Ala LpxD was stored in 10 mM Tris-HCl, pH 7.5, 500 mM NaCl and 1 mM TCEP, and Met290Cys LpxD was purified and stored in the absence of a reductant.

***In vitro* assay of LpxD.** The LpxD catalysed conversion of [α-^{32}P]UDP-3-acylGlcN to [α-^{32}P]UDP-2,3-diacylGlcN was monitored by thin-layer chromatography as previously reported[25]. The 20 µl assay mixture containing 40 mM HEPES (pH 7.4), 0.02 mg ml^{-1} BSA, 1.4 nM pure wild-type *E. coli* LpxD (or 4.2 nM Met290Cys LpxD), 6 µM *R*-3-hydroxymyristoyl-ACP (β-OH-C14-ACP) and 4 µM [α-^{32}P]UDP-3-*O*-(β-OH-C14)-GlcN (0.005–0.04 µCi $µl^{-1}$) was equilibrated at 30 °C, and the reaction was initiated by the addition of the enzyme. DTT (100 mM) was added to assess its effect on the catalysis of the wild-type and mutant LpxD. A 1 µl aliquot of the reaction mixture was spotted onto a silica gel 60 plate (EMD Chemicals) at various time-points. After drying under a cold air stream, plates were developed with a chloroform/methanol/water/acetic acid solvent (25:15:4:2, v:v:v:v). The plates were dried and exposed overnight to a Molecular Dynamics PhosphorImager Screen. The conversion rate of the lipid substrate to product was measured using ImageQuant software.

Expression and purification of holo-ACP. The *acp* gene was amplified from *E. coli* W3110A genomic DNA using the ACP-forward and ACP-reverse primers (Supplementary Table 1) engineered to introduce downstream of the ACP coding region a spacer sequence, a ribosome binding site, followed by a translation spacer element (denoted PCR product A). The ACPS-forward and ACPS-reverse primers (Supplementary Table 1) were used to amplify the *acps* gene from W3110A genomic DNA, which contained a spacer sequence, a ribosome binding site and a translation spacer element upstream of the ACPS coding region (denoted PCR product B). The ACP-forward and ACPS-reverse primers were subsequently used to generate PCR product C, which contained *acp* and *acps* genes separated by the spacer sequence, ribosome binding site and the translation spacer element. PCR product C was cloned into a pET16b* vector, which contained an engineered tobacco etch virus (TEV) protease cleavage site instead of Factor Xa, using the NdeI and XhoI restriction sites. *E. coli* DH5α-competent cells were subsequently transformed with plasmid and transformants were selected at 37 °C on a LB-agar plate supplemented with 100 µg ml^{-1} ampicillin. The sequence of the plasmid encoding the N-terminally histidine-tagged ACP and non-tagged ACPS (pET16b*-AM) was confirmed by DNA sequencing.

ACP and ACPS were co-expressed using the pET16b*-AM vector in C41(DE3) *E. coli* cells and cultured at 37 °C in 1 litre of LB broth supplemented with 100 µg ml^{-1} ampicillin. Upon reaching an absorbance at 600 nm of 0.6, the expression was induced for 6 h with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at 30 °C. All the subsequent procedures were performed at 4 °C. The cells were collected, washed, re-suspended and lysed in 20 mM HEPES, pH 8.0 containing 10% glycerol, 200 mM NaCl and 2 mM DTT (buffer A), supplemented with 20 mM imidazole. After removal of cell debris by centrifugation at 100,000*g* for 1 h, the soluble fraction was loaded onto a Ni-NTA (Qiagen) column equilibrated with buffer A and 20 mM imidazole. The Ni-NTA column was washed with 10 column-volumes and His_{10}-ACP was eluted in one step by buffer A supplemented with 250 mM imidazole over the equivalent of five column-volumes. Although ACPS was not histidine-tagged, it co-purified with His_{10}-ACP. The elution fraction was allowed to slowly rock with 1 mg of TEV protease and 2 mM EDTA overnight. Subsequently, the protein mixture was dialysed against 20 mM MES, pH 6.0 and 2 mM DTT overnight. The TEV protease and residual His_{10}-ACP were removed by passing the mixture through a 2 ml Ni-NTA column. The resulting flow-through fraction containing ACP was loaded onto a 5 ml QFF ion exchange column (GE). ACP was eluted separately from ACPS by applying a linear gradient of 20 mM MES, pH 6.0, 0–500 mM NaCl and 2 mM DTT over 50 column-volumes. The fractions corresponding to ACP were pooled together and dialysed against 100 mM Tris-HCl, pH 7.1 and 200 mM NaCl (buffer B) overnight. Holo-ACP was separated from residual apo-ACP by mixing the ACP sample with 5 ml of ThioPropyl Sepharose 6b resin (Sigma-Aldrich) overnight. Apo-ACP, which lacks any free thiol group, did not bind to the resin, whereas holo-ACP covalently attached to the resin by the terminal thiol group of its phosphopantetheine moiety. After washing the Thiopropyl Sepharose 6b with 25 ml of buffer B, holo-ACP was eluted in 25 ml of buffer B supplemented with 25 mM DTT. The complete removal of apo-ACP was confirmed by electrospray ionization mass spectrometry. The elution fraction was concentrated to approximately 4–10 ml and loaded onto a High Load 26/60 Superdex 200 gel filtration column equilibrated with 10 mM Tris-HCl, pH7.5, 200 mM NaCl and 2 mM DTT. The relevant eluted fractions were concentrated to approximately 20 mg ml^{-1} and stored in aliquots at −80 °C.

Production of β-OH-C14-ACP. Holo-ACP was charged enzymatically with *R*-3-hydroxymyristic acid (Santa Cruz Biotechnology) by soluble acyl-ACP synthetase from *V. harveyi*[30]. The soluble acyl-ACP synthetase (*aasS*) gene was synthesized by GenScript. The *aasS* gene was subcloned into pET-16b expression vector, over-expressed, and the His_6-AasS was purified using nickel affinity chromatography as reported by Jiang *et al.*[30]. To generate β-OH-C14-ACP, 0.1 mM of holo-ACP was mixed with 0.001 mM AasS and 0.3 mM of the fatty acid (β-OH-C14) at room temperature for 3 h in a buffer containing 100 mM Tris-HCl, pH 7.8, 10 mM ATP and 10 mM $MgCl_2$. To separate acyl-ACP from AasS, the reaction mixture was loaded onto a High Load 26/60 Superdex 200 gel filtration column equilibrated with 10 mM Tris-HCl, pH 7.5 and 200 mM NaCl. The complete conversion of holo-ACP to β-OH-C14-ACP was confirmed both by electrospray ionization mass spectrometry and 2.5 M urea (19%) polyacrylamide (pH 9.5) gel electrophoresis[31].

Crystallization and structure determination. Before crystallization, either acyl-ACP or holo-ACP was mixed with the wild-type or the catalytically inactive His239Ala LpxD to preform the protein–protein complex. Crystals of holo-ACP–LpxD were grown at 15 °C by mixing the protein solution with the precipitant (0.1 M MES pH 6.0, 0.2 M lithium sulphate, 20% PEG 4000) in ratios of 1:1 and 1:1.5. Crystals achieved full size in approximately 45 days. The crystals were transferred to a cryo-solution using a 50:50 ratio of paratone to mineral oil and immediately cryo-cooled to −180 °C in liquid nitrogen. Hydrolysed-acyl-ACP–LpxD crystals were obtained by equilibrating the protein mixture against a well solution containing 0.1 M MES pH 6.5, 0.2 M ammonium sulphate, 20% PEG 8000 and incubating at 15 °C. The crystals were collected on day 10 and cryo-cooled using a solution containing ammonium sulphate, 33% PEG 8000, 5 mM Tris-HCl pH 7.5, 190 mM NaCl and 20% of the cryoprotectant ethylene glycol. To trap the intact-acyl-ACP–LpxD complex the His239Ala-LpxD mutant was purified in the presence of the reducing agent TCEP instead of DTT in an attempt to reduce hydrolysis of the thioester bond. His239Ala-LpxD and acyl-ACP were pre-mixed using a 1:1 molar ratio before crystallization trials. The well solution and the cryoprotectant were the same as that described for the hydrolysed-acyl-ACP structure; however, a ratio of 1.5:1 of protein to well solution was required.

The diffraction data were collected on the SER-CAT 22-BM and 22-ID beamlines at the Advanced Photon Source at Argonne National Laboratory. Data were processed using the HKL2000[32] software suite (Supplementary Table 2). Although the data for both intact- and holo-ACP complexes were processed to 2.1 and 2.13 Å, respectively, the hydrolysed-acyl-ACP complex was trimmed to 2.9 Å owing to the data completeness being unsatisfactory in higher resolution bins (that is, well below 70%). This resulted in a higher signal-to-noise ratio in the 2.9 Å resolution bin as well as a much lower R_{merge} value. The crystals of each complex belong to the *P*1 space group. The structures were solved by the molecular replacement method using the program PHASER in the PHENIX software suite[33,34] and the previously determined *E. coli* LpxD structure (Protein Data Bank 3EH0)[19] as the search model. Two trimers of LpxD were observed in the triclinic unit cell for the intact-acyl-ACP, hydrolysed-acyl-ACP and holo-ACP co-crystal structures. In each case, however, ACP was intentionally omitted during the molecular replacement process; instead, it was manually rebuilt into unbiased, contiguous $F_o - F_c$ difference electron density by first rigid-fitting the *E. coli apo*-ACP coordinates (Protein Data Bank 1T8K)[35] into the resulting maps. The models were rebuilt using COOT[36] and iterative structure refinement with restrained and TLS options was performed using PHENIX[33]. For the lower-resolution hydrolysed-acyl-ACP structure, hydrogens were included for refinement with automated optimization of X-ray/stereochemistry and ADP weights selected. Additionally, we used the high-resolution *E. coli apo*-ACP structure (Protein Data Bank 1T8K) as a reference model. The molecular coordinates and restraints of the 4′-PPT, β-OH-C14-4′-PPT and free β-OH-C14 fatty-acid ligands were generated by using either the Dundee PRODRG2 Server[37] or PHENIX Elbow[33]. Composite omit map and simulated annealing omit map calculations were conducted using CNS[38]. The protein–ligand interactions were identified by AREAIMOL calculations in the CCP4 suite[39]. The quality of the final models was validated using MOLPROBITY[40]. The data statistics are reported in Supplementary Table 2. Molecular figures were generated using PyMOL[41].

A total of six ACP molecules were present in the intact-acyl-ACP structure (Supplementary Fig. 3). Both the hydrolysed- and holo-ACP structures showed partly bound states of ACP within two or three LpxD active sites as interpreted by the lack of contiguous electron density for most of the ACP backbone. As a result, only a portion of ACP that includes Ser 36 and its 4′-PPT prosthetic group was included in the final model: (1) hydrolysed-acyl-ACP, residues 35–44 of chain L, residues 6–15, 27–53 and 62–73 of chain I, residues 1–15 and 27–73 of chain G; (2) holo-ACP, residues 35–44 of chain H and chain L. The remaining LpxD active sites in the hydrolysed- and holo-ACP structures contain fully modelled ACPs. In all three co-crystal structures, the electron density maps indicated that the N-terminal methionine of ACP was present and forms a key lattice contact with neighbouring molecules of LpxD. Two more residues (Ser–His) in the holo-ACP structure that remain from the TEV cleavage site could also be modelled.

The observed electron density for fully modelled intact-, hydrolysed- and holo-ACPs were weaker in some regions, especially on the backside of the molecule which faces solvent; however, the placement of side-chains was apparent at the protein–protein interface (Supplementary Fig. 3). This implicates conformational heterogeneity throughout the lattice. Nonetheless, in all three co-crystal structures, electron density was apparent in every active site for all atoms of the 4′-PPT group. In both the intact- and hydrolysed-forms, pronounced electron density was also present for the β-OH-C14 acyl chains located in the *N*-channel. In addition, electron density indicated that the hydrolysed-acyl-ACP complex included two extra molecules of β-OH-C14 fatty acid bound to the hydrophobic *O*-channel, although the density was weaker towards the terminal carbon atoms of the acyl chains. To investigate the origin of this second fatty acid, we mixed acyl-ACP (12.86 mg ml^{-1}) in a 1:2.25 v:v ratio with wild-type LpxD (26.38 mg ml^{-1}) pre-incubated with 1 mM DTT in a solution that was consistent with the condition used for crystallization. The protein solution was incubated at 15 °C, and aliquots were taken at different time points and stored at −80 °C. Samples were run on a 2.5 M urea (19%) polyacrylamide (pH 9.5) gel[31], which showed that both DTT (a known phenomenon[42]) and LpxD enhance the cleavage of the thioester bond of acyl-ACP (Supplementary Fig. 10). This observation most likely explains why free β-OH-C14 fatty acid was available to bind the *O*-channel of LpxD.

30. Jiang, Y., Chan, C. H. & Cronan, J. E. The soluble acyl-acyl carrier protein synthetase of *Vibrio harveyi* B392 is a member of the medium chain acyl-CoA synthetase family. *Biochemistry* **45,** 10008–10019 (2006).
31. Rock, C. O., Cronan, J. E. Jr & Armitage, I. M. Molecular properties of acyl carrier protein derivatives. *J. Biol. Chem.* **256,** 2669–2674 (1981).
32. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
33. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
35. Qiu, X. & Janson, C. A. Structure of apo acyl carrier protein and a proposal to engineer protein crystallization through metal ions. *Acta Crystallogr. D* **60,** 1545–1554 (2004).
36. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
37. Schuttelkopf, A. W. & van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60,** 1355–1363 (2004).
38. Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
39. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
40. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
41. DeLano, W. L. The PyMOL Molecular Graphics System v. 1.3r1 (Schrödinger, LLC, New York, New York, 2012).
42. Stokes, G. B. & Stumpf, P. K. Fat metabolism in higher plants. The nonenzymatic acylation of dithiothreitol by acyl coenzyme A. *Arch. Biochem. Biophys.* **162,** 638–648 (1974).

LETTER

doi:10.1038/nature12810

Trapping the dynamic acyl carrier protein in fatty acid biosynthesis

Chi Nguyen[1]*, Robert W. Haushalter[2]*, D. John Lee[2]*, Phineus R. L. Markwick[2,3,4], Joel Bruegger[1], Grace Caldara-Festin[1], Kara Finzel[2], David R. Jackson[1], Fumihiro Ishikawa[2], Bing O'Dowd[2], J. Andrew McCammon[2,4], Stanley J. Opella[2], Shiou-Chuan Tsai[1] & Michael D. Burkart[2]

Acyl carrier protein (ACP) transports the growing fatty acid chain between enzymatic domains of fatty acid synthase (FAS) during biosynthesis[1]. Because FAS enzymes operate on ACP-bound acyl groups, ACP must stabilize and transport the growing lipid chain[2]. ACPs have a central role in transporting starting materials and intermediates throughout the fatty acid biosynthetic pathway[3–5]. The transient nature of ACP–enzyme interactions impose major obstacles to obtaining high-resolution structural information about fatty acid biosynthesis, and a new strategy is required to study protein–protein interactions effectively. Here we describe the application of a mechanism-based probe that allows active site-selective covalent crosslinking of AcpP to FabA, the *Escherichia coli* ACP and fatty acid 3-hydroxyacyl-ACP dehydratase, respectively. We report the 1.9 Å crystal structure of the crosslinked AcpP–FabA complex as a homodimer in which AcpP exhibits two different conformations, representing probable snapshots of ACP in action: the 4′-phosphopantetheine group of AcpP first binds an arginine-rich groove of FabA, then an AcpP helical conformational change locks AcpP and FabA in place. Residues at the interface of AcpP and FabA are identified and validated by solution nuclear magnetic resonance techniques, including chemical shift perturbations and residual dipolar coupling measurements. These not only support our interpretation of the crystal structures but also provide an animated view of ACP in action during fatty acid dehydration. These techniques, in combination with molecular dynamics simulations, show for the first time that FabA extrudes the sequestered acyl chain from the ACP binding pocket before dehydration by repositioning helix III. Extensive sequence conservation among carrier proteins suggests that the mechanistic insights gleaned from our studies may be broadly applicable to fatty acid, polyketide and non-ribosomal biosynthesis. Here the foundation is laid for defining the dynamic action of carrier-protein activity in primary and secondary metabolism, providing insight into pathways that can have major roles in the treatment of cancer, obesity and infectious disease.

In *E. coli*, AcpP interacts with at least 12 enzymes involved in fatty acid biosynthesis, plus nine other enzymes from disparate biosynthetic pathways (Fig. 1a and Supplementary Fig. 1)[6–10]. AcpP sequesters growing metabolites in an interior hydrophobic cavity that protects these intermediates from non-selective reactivity[11], and selective protein–protein interactions are believed to be a prerequisite for the delivery of ACP-bound substrate to its catalytic partners[3]. Given the importance of ACP–protein interactions in metabolism and cell regulatory processes, understanding this 'switchblade mechanism' (Fig. 1c) is crucial[12], although this has proven elusive owing to the inherently transient nature of ACP–partner complexes[13].

We recently deployed synthetic probes to study ACP activity and protein–protein interactions[14], including a sulphonyl-3-alkyne-based probe (**1**, Fig. 1b) designed to capture ACP in functional association with 3-hydroxyacyl-ACP dehydratase with demonstrated specificity between *E. coli* AcpP and FabA (Fig. 1c, d)[15,16]. Probe **1** applied to AcpP and FabA creates a uniformly crosslinked species (AcpP–FabA) that forms reproducible crystals in tag-free form (Supplementary Fig. 4). No AcpP–FabA protein complex crystals form without **1**, demonstrating the necessity of applying probes such as **1** to capture ACP in action.

The AcpP–FabA crystals diffracted to 1.9 Å (Supplementary Table 2), and we solved the AcpP–FabA crystal structure by molecular replacement using the apo-FabA dimer (Protein Data Bank (PDB) code 1MKA)[17] and two butyryl-AcpP (PDB 2K94) as search templates (Supplementary Fig. 5). Final refinement revealed the structure of a one-to-one ratio between AcpP and FabA yielding a $AcpP_2$–$FabA_2$ complex (Fig. 2a), consistent with protein-sizing data in solution. The dimeric FabA forms a 'double hotdog' topology[18], with two antiparallel 'hotdog' helices surrounded by a combined 14-stranded β-sheet (Fig. 2a)[17,19]. The two AcpP monomers adopt a four-helix bundle fold[3] and dock helices II–III with the β5–6 loop of FabA (Fig. 2d). The contact area is small (503 $Å^2$ and 539 $Å^2$ for the first and second AcpP–FabA interface), consistent with the transient nature of AcpP–partner interactions. AcpP has a negatively charged cleft between helices II and III, which interacts with a positively charged arginine-rich patch on FabA, the 'Positive Patch', Fig. 2c)[3]. The AcpP–FabA interface interactions are primarily electrostatic but also include conserved, hydrophobic residues (Fig. 2d, detailed in Supplementary Information). The high sequence conservation of negatively charged residues on helices II and III of AcpP at the AcpP–FabA interface (Supplementary Figs 2, 3) is consistent with previous reports of ACP–partner complex structures (Fig. 4d)[7,20–23], strongly supporting the presence of the 'Positive Patch' in ACP partner proteins.

Only the position of R137 differs between the two FabA protomers, but in the ACP structures many residues of helix III move extensively (Fig. 2d and Supplementary Information), resulting in different topology near the contact interface between helices II and III (Fig. 2c and Supplementary Information). Thus the second AcpP–FabA interaction probably represents a snapshot when AcpP completes its docking with FabA, resulting in less disorder of AcpP2. Accordingly, the first AcpP–FabA interaction would represent a snapshot of AcpP in transition, where extensive movement of helix III is necessary in order for AcpP1 to bind or dissociate from FabA.

The natural FAS substrates contain both 4′-phosphopantetheine (PPant) and acyl-chain moieties (Fig. 1b), and the application of probe **1** shows how both bind to FabA (Supplementary Figs 6, 7). Probe **1** covalently connects the active site S36 of AcpP and H70 of FabA and binds in a highly conserved tunnel of FabA (detailed in Supplementary Information). Unlike acyl–AcpP structures that contain a hydrophobic interior pocket to sequester the acyl chain[13], the AcpP in the AcpP–FabA complex structure contains no interior pocket (Fig. 2e) and closely resembles apo-AcpP (Supplementary Table 4)[24], because five conserved hydrophobic

[1]Departments of Molecular Biology and Biochemistry, Chemistry, and Pharmaceutical Sciences, University of California, Irvine, California 92697, USA. [2]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, USA. [3]San Diego Supercomputer Center, La Jolla, California 92093, USA. [4]Howard Hughes Medical Institute, La Jolla, California 92093, USA.
*These authors contributed equally to this work.

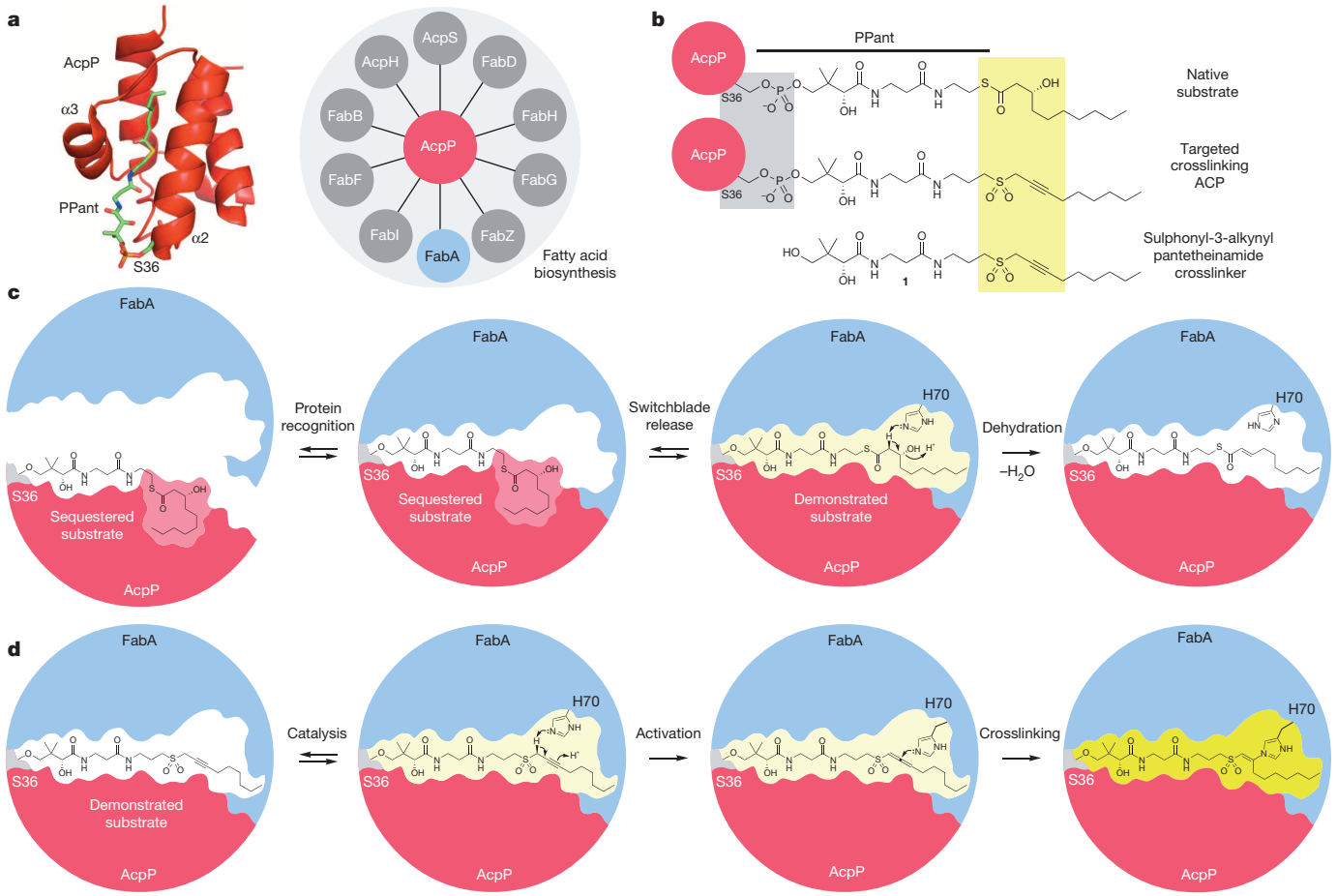


Figure 1 | *E. coli* AcpP and crosslinking strategy. **a**, AcpP is a small, acidic protein comprised of four α-helices that interacts with at least 21 catalytic enzymes, 12 of which belong to FAS (10 shown here). The apolar interior of helix II (α2) and helix III (α3) form a hydrophobic cavity that sequesters the growing metabolite attached to the PPant arm. **b**, A native substrate of FabA (top) and modified AcpP with targeted sulphonyl-3-alkynyl crosslinking probe (middle), derived from the crosslinking pantetheinamide analogue **1** (bottom). **c**, Proposed mechanism of FabA. Protein–protein interactions between AcpP and FabA induce release of the sequestered substrate from AcpP into the active site of FabA, where dehydration is catalysed. **d**, Crosslinking strategy to form AcpP–FabA with mechanism-based crosslinking probe **1**.

residues between helices II and III move inward and collapse the interior pocket (Fig. 2f). This marked change reflects a dynamic AcpP moving from the sequestered-substrate state to the open switchblade state to position the substrate within the partner enzyme FabA (Fig. 1c, d).

Further characterization of the dynamic interactions between AcpP and FabA was achieved through comparisons with two-dimensional $^{1}H/^{15}N$ heteronuclear single quantum coherence (HSQC) spectra of holo-AcpP and octanoyl-AcpP. Many resonances displayed chemical shift perturbations (CSPs) in residues that define the hydrophobic pocket of AcpP (Supplementary Fig. 11). Titrating unlabelled apo-FabA into each sample allowed us to observe CSPs resulting from dynamic, non-covalent association with FabA in solution (Fig. 3a, b). In the holo-AcpP–FabA titration experiment, we observed significant CSPs in residues spanning helices II and III and the adjacent loops (Supplementary Fig. 12). In the octanoyl-AcpP–FabA titration experiment, additional CSPs were observed in residues lining the hydrophobic pocket of AcpP (Fig. 3c), which we attribute to the translocation of the bound acyl chain out of the AcpP pocket into the FabA active site to complete the 'switchblade' process.

We then acquired a two-dimensional HSQC spectrum of the AcpP–FabA complex in 1:2 stoichiometry, with one AcpP crosslinked to each FabA homodimer. When overlaying the HSQC spectra of the AcpP–FabA complex with the holo- and octanoyl-AcpP titration data (Fig. 3a, b) obtained with experiments incorporating transverse relaxation optimized spectroscopy[25], we observed a striking correlation between the CSP shifts in the two AcpP species and the HSQC spectrum of the complex; the chemical shift of each residue migrates towards the observed chemical shift in the crosslinked complex as the concentration of FabA increases. The similarities between CSPs of the transient binding event and our crosslinked complex (Fig. 3c, f) indicate that the binding conformation in the AcpP–FabA complex is truly indicative of the natively bound conformation.

From the crystal structure we identified acidic residues E41, E47, E53 and E60 of AcpP that interact with FabA (Fig. 2d and Supplementary Fig. 6), and correspondingly observed large CSPs in these residues between helices II and III (Fig. 3c–e). In addition, we found large CSPs in the hydrophobic helix II residues observed in the binding interface of the crystal structure, such as L37, V40 and M44 (Figs 2d and 3 and Supplementary Fig. 6). Similarly, crystal structure observations are consistent with CSP plots for both octanoyl-AcpP with crosslinked AcpP–FabA (Fig. 3c–e and Supplementary Table 8): the strong CSPs for A59, E60, E41 and E47 correspond with side-chain interactions within the Positive Patch of FabA. Large CSPs in S36, L37 and D38 correspond with a change in the PPant position as it extends into the FabA. T63 undergoes considerable rotation when comparing the octanoyl-AcpP and the AcpP–FabA complex. Ultimately, these CSP observations both complement and corroborate binding observations found in the crystal structure.

To study the detailed dynamics of AcpP and its interaction with FabA, we measured residual dipolar couplings (RDCs)[26] from weakly

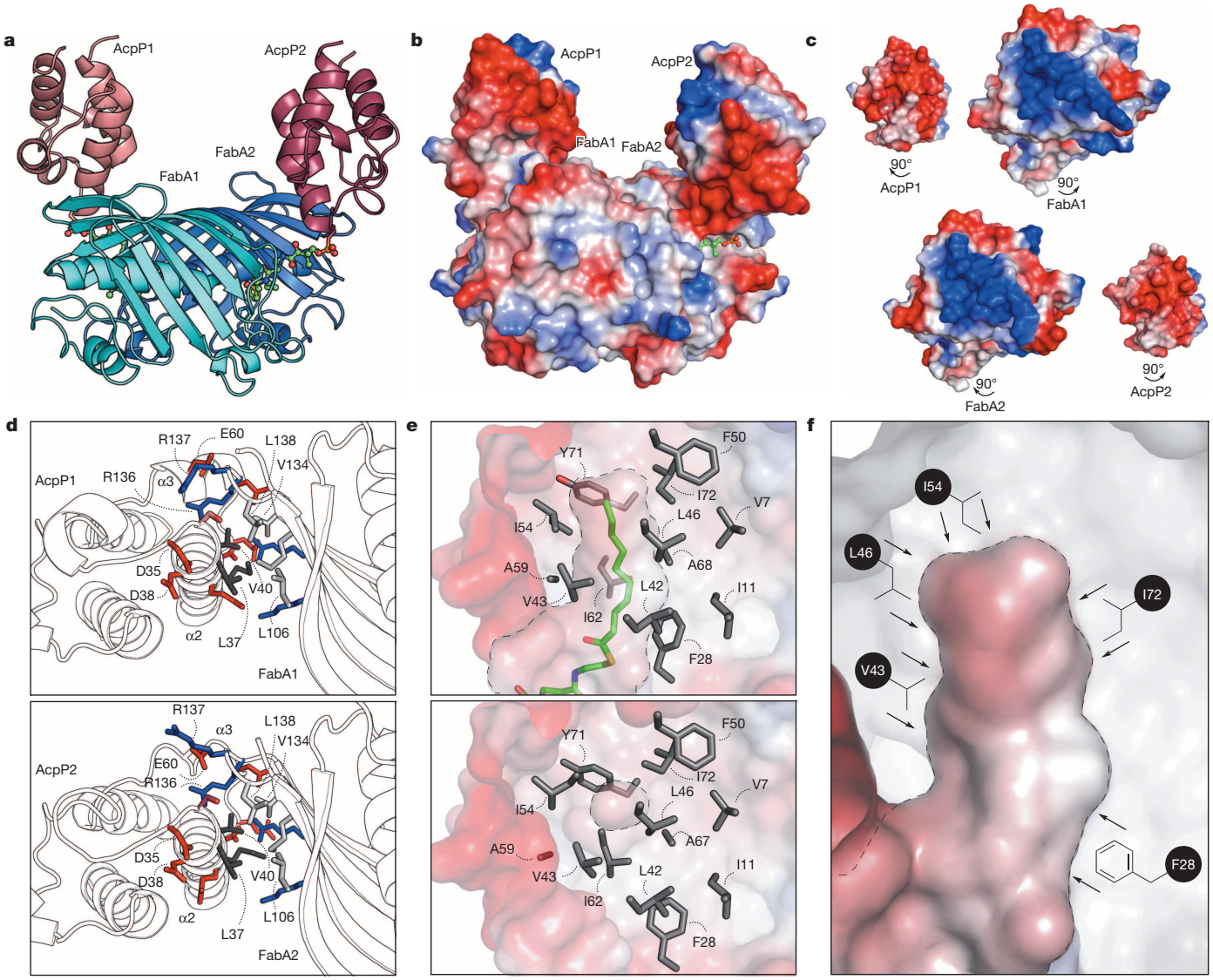


Figure 2 | Structure of crosslinked AcpP–FabA. **a**, X-ray crystal structure of AcpP–FabA at 1.9 Å. **b**, The molecular surface mapped with calculated vacuum electrostatic potential of AcpP–FabA. Blue shading indicates electropositive and red shading indicates electronegative protein surfaces. **c**, Rotation of **b** by 90° at the interfaces between each AcpP–FabA to visualize electrostatic pairing. **d**, Expanded view of both interfaces in AcpP–FabA, indicating salt bridges and hydrophobic interactions between helix II (α2) and helix III (α3) of AcpP and the Positive Patch of FabA. **e**, Comparison between hydrophobic cleft of AcpP with sequestered substrate (top, from PDB 2FAE, with long interior hydrophobic cavity outlined with dashed line) and AcpP1 in AcpP–FabA (bottom, reduced interior cavity). **f**, The interior cavity of 2FAE labelled with the hydrophobic residues. The contraction of these hydrophobic residues collapses the interior cavity in AcpP–FabA.

aligned samples of octanoyl-AcpP and AcpP in the AcpP–FabA complex (Fig. 4a, b). The empirical RDCs were combined with accelerated molecular dynamics simulations[27] to provide structural-dynamic information in the slow RDC-optimized (microsecond) regime (detailed in Supplementary Information)[28]. Within this framework, we identified the optimal acceleration parameters, and hence optimal conformational space sampling criteria for the correlation of experimental and theoretical RDCs (Fig. 4a), and calculated the averaged NH order parameters at both fast (nanosecond) and slow (microsecond) timescales (Fig. 4b). On nanosecond timescales, no substantial differences in the order parameters between octanoyl-AcpP and AcpP–FabA are observed. By contrast, in the microsecond time regime, substantial differences in the structural-dynamic behaviour of octanoyl-AcpP and AcpP–FabA are identified, indicating that AcpP is markedly stabilized in the presence of FabA, especially in the amino-terminal region of helix II and the helix II–III loop. There is a notable correlation between the AcpP–FabA binding interface observed in the crystal structure (Fig. 2), the NMR titration data (Fig. 3) and molecular dynamics simulations (Fig. 4b, c), all highlighting key dynamic residues.

These results provide a window into the dynamic properties of AcpP, which sequesters elongated substrates in its interior cavity with motion at the helix II–III loop on a microsecond timescale. A probable order of events is that the Positive Patch of FabA first interacts with PPant attached to S36 of AcpP. Once in proximity, residues R132 and K161 of FabA form salt bridges with E41 and E47 on helix II of AcpP, anchoring the complex, whereas R136 and R137 serve to pry away helix III of AcpP through interactions with A59 and E60, thus disrupting shielding of the sequestered substrate. V40 and L37 on AcpP form hydrophobic interactions with L138 and V134 on FabA. All of these binding events serve to stabilize AcpP in an open conformation, allowing the sequestered substrate to release from AcpP and insert into the pocket of FabA as the AcpP hydrophobic cavity collapses. Together, these results provide an unprecedented verification of the switchblade mechanism (Fig. 1c). We surmise that the identity of the sequestered substrate can affect the

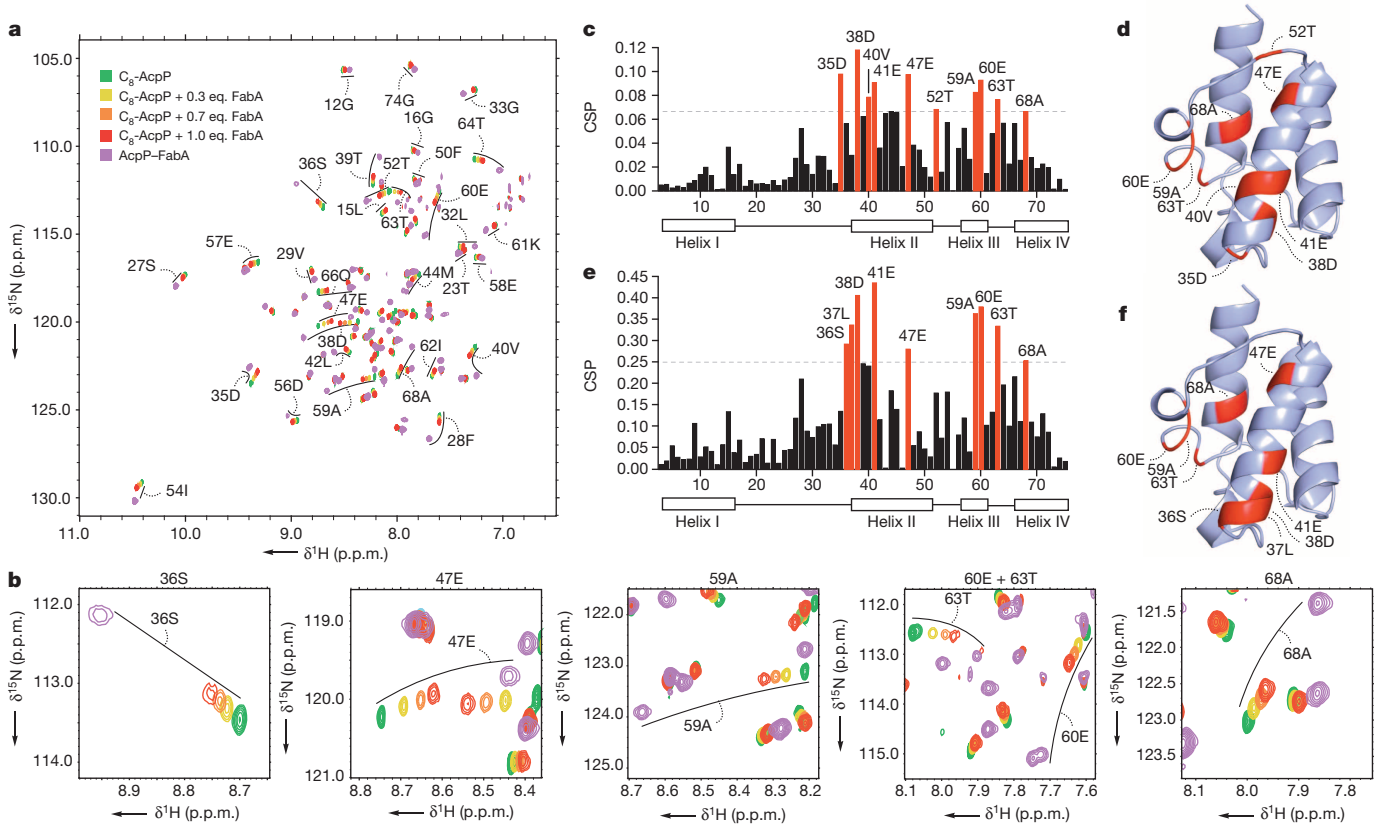


Figure 3 | NMR studies. **a**, HSQC titration spectra of ^{15}N-octanoyl-AcpP in the absence of FabA (green), and with increasing (yellow to red) concentrations of FabA. CSPs are observed in AcpP residues that interact with FabA or the bound acyl chain. In magenta is the overlaid HSQC of crosslinked ^{15}N-AcpP–FabA. **b**, Expanded views of select residues. **c**, CSPs were measured for each ^{15}N-octanoyl-AcpP residue in the absence and presence of 1 molar equivalent of FabA and plotted by residue number. **d**, AcpP residues from **c**, where CSPs larger than 0.065 parts per million (p.p.m.) are indicated in red. **e**, CSPs measured between ^{15}N-octanoyl-AcpP and ^{15}N-AcpP–FabA were measured and plotted by residue number. **f**, AcpP residues from **e**, where CSPs larger than 0.25 p.p.m. are indicated in red. In NMR convention, protein residue number precedes residue letter; the converse applies with crystallography.

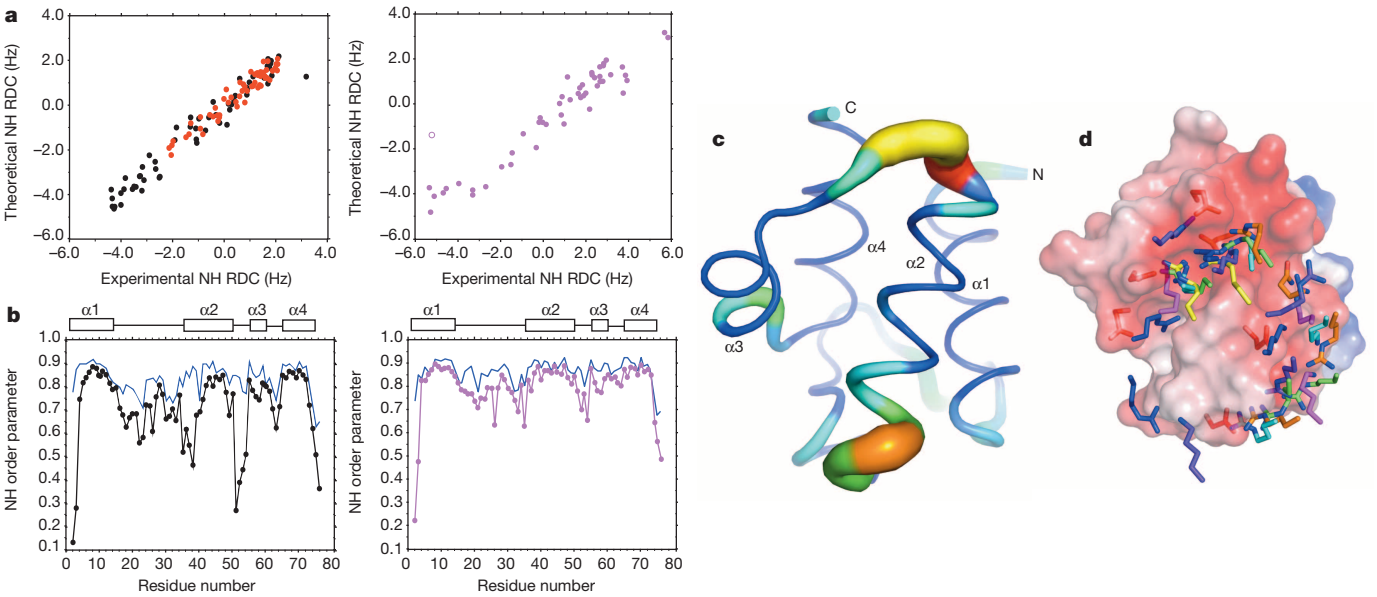


Figure 4 | Molecular dynamics and protein–protein interactions. **a**, Experimental RDC data correlated with theoretical RDCs. ^{15}N-octanoyl-AcpP with Pf1 bacteriophage (left, black) and 5% neutral charge compressed polyacrylamide gel (left, red), and crosslinked AcpP–FabA with Pf1 bacteriophage (right, magenta). **b**, Order parameter calculations of octanoyl-AcpP (left) and AcpP–FabA (right). Nanosecond (blue) timescale compared to microsecond (RDC-optimized) (dotted) timescale routines. **c**, Sausage plot of order parameter differences on the microsecond timescale between octanoyl-AcpP and AcpP–FabA. Colour and thickness depict relative disorder, where red represents maximal difference of 0.5 (detailed in Supplementary Information). **d**, Residues of the Positive Patch mediating protein–protein interactions in known structures. Blue, FabA; cyan, ACP–ACP synthase (PDB: 4DXE); purple, ACP–stearoyl-ACP desaturase (PDB 2XZ0); orange, holo ACP–ACP synthase (PDB 1F80); magenta, ACP–P450 (PDB 3EJB); yellow, ACP–STAS (PDB 3NY7); and green, ACP–BioH (PDB 4ETW).

positioning of helix II and helix III of AcpP, thereby modulating successful binding and switchblade events for selective catalysis.

The application of crosslinking probes to gain structural insights now lays a foundation for defining the dynamic events associated with the mechanism of action of ACP. The approach can be applied to other carrier protein partners from primary and secondary metabolism such as FAS, polyketide synthase and non-ribosomal peptide synthetase[3,29,30], as well as other carrier-protein-dependent pathways that have major roles in the treatment of cancer, obesity and infectious disease (Supplementary Table 9).

METHODS SUMMARY

All proteins used were overproduced in *E. coli* BL21(DE3) (Novagen) and purified by Ni-affinity followed by fast protein liquid chromatography. The AcpP–FabA complex was generated as reported previously and crystallized at room temperature (25 °C) by sitting drop vapour diffusion at 30 mg ml^{-1} in 10 mM sodium phosphate (pH 8.0), 350 mM sodium acetate, 1 M LiCl and 35% PEG3350. Data were collected on beamline 12-2 at the Stanford Synchrotron Radiation Lightsource and beamline 8.2.2 at the Advanced Light Source and processed with HKL2000. The AcpP–FabA crystallographic phases were determined by molecular replacement using FabA and AcpP as the search template. Protein NMR data were collected at the University of California, San Diego Biomolecular NMR facility. Details of the molecular dynamics simulations are included in Supplementary Discussion. Detailed experimental procedures are described in the Supplementary Methods.

Received 15 May 2013; accepted 23 October 2013.
Published online 22 December 2013.

1. Chan, D. I. & Vogel, H. J. Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430,** 1–19 (2010).
2. Rock, C. O. & Cronan, J. E., Jr. Acyl carrier protein from *Escherichia coli. Methods Enzymol.* **71,** 341–351 (1981).
3. Crosby, J. & Crump, M. P. The structural role of the carrier protein—active controller or passive carrier. *Nat. Prod. Rep.* **29,** 1111–1137 (2012).
4. Magnuson, K., Jackowski, S., Rock, C. O. & Cronan, J. E., Jr. Regulation of fatty acid biosynthesis in *Escherichia coli. Microbiol. Rev.* **57,** 522–542 (1993).
5. Joshi, A. K., Witkowski, A., Berman, H. A., Zhang, L. & Smith, S. Effect of modification of the length and flexibility of the acyl carrier protein–thioesterase interdomain linker on functionality of the animal fatty acid synthase. *Biochemistry* **44,** 4100–4107 (2005).
6. Issartel, J. P., Koronakis, V. & Hughes, C. Activation of *Escherichia coli* prohaemolysin to the mature toxin by acyl carrier protein-dependent fatty acylation. *Nature* **351,** 759–761 (1991).
7. Agarwal, V., Lin, S., Lukk, T., Nair, S. K. & Cronan, J. E., Jr. Structure of the enzyme-acyl carrier protein (ACP) substrate gatekeeper complex required for biotin synthesis. *Proc. Natl Acad. Sci. USA* **109,** 17406–17411 (2012).
8. Anderson, M. S., Bulawa, C. E. & Raetz, C. R. The biosynthesis of gram-negative endotoxin. Formation of lipid A precursors from UDP-GlcNAc in extracts of *Escherichia coli. J. Biol. Chem.* **260,** 15536–15541 (1985).
9. Jordan, S. W. & Cronan, J. E., Jr. A new metabolic link. The acyl carrier protein of lipid synthesis donates lipoic acid to the pyruvate dehydrogenase complex in *Escherichia coli* and mitochondria. *J. Biol. Chem.* **272,** 17903–17906 (1997).
10. Lu, Y. J. *et al.* Acyl-phosphates initiate membrane phospholipid synthesis in Gram-positive pathogens. *Mol. Cell* **23,** 765–772 (2006).
11. Roujeinikova, A. *et al.* Crystallization and preliminary X-ray crystallographic studies on acyl-(acyl carrier protein) from *Escherichia coli. Acta Crystallogr. D* **58,** 330–332 (2002).
12. Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316,** 288–290 (2007).
13. Roujeinikova, A. *et al.* Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J. Mol. Biol.* **365,** 135–145 (2007).
14. Meier, J. L. & Burkart, M. D. The chemical biology of modular biosynthetic enzymes. *Chem. Soc. Rev.* **38,** 2012–2045 (2009).
15. Ishikawa, F., Haushalter, R. W. & Burkart, M. D. Dehydratase-specific probes for fatty acid and polyketide synthases. *J. Am. Chem. Soc.* **134,** 769–772 (2012).
16. Endo, K., Helmkamp, G. M. Jr & Bloch, K. Mode of inhibition of β-hydroxydecanoyl thioester dehydrase by 3-decynoyl-*N*-acetylcysteamine. *J. Biol. Chem.* **245,** 4293–4296 (1970).
17. Leesong, M., Henderson, B. S., Gillig, J. R., Schwab, J. M. & Smith, J. L. Structure of a dehydratase–isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. *Structure* **4,** 253–264 (1996).
18. Zhuang, Z. *et al.* Divergence of function in the hot dog fold enzyme superfamily: the bacterial thioesterase YciA. *Biochemistry* **47,** 2789–2796 (2008).
19. Moynié, L. *et al.* Structural insights into the mechanism and inhibition of the β-hydroxydecanoyl-acyl carrier protein dehydratase from *Pseudomonas aeruginosa*. *J. Mol. Biol.* **425,** 365–377 (2013).
20. Guy, J. E. *et al.* Remote control of regioselectivity in acyl-acyl carrier protein-desaturases. *Proc. Natl Acad. Sci. USA* **108,** 16594–16599 (2011).
21. Parris, K. D. *et al.* Crystal structures of substrate binding to *Bacillus subtilis* holo-(acyl carrier protein) synthase reveal a novel trimeric arrangement of molecules resulting in three active sites. *Structure* **8,** 883–895 (2000).
22. Cryle, M. J. & Schlichting, I. Structural insights from a P450 Carrier Protein complex reveal how specificity is achieved in the P450(BioI) ACP complex. *Proc. Natl Acad. Sci. USA* **105,** 15696–15701 (2008).
23. Babu, M. *et al.* Structure of a SLC26 anion transporter STAS domain in complex with acyl carrier protein: implications for *E. coli* YchM in fatty acid metabolism. *Structure* **18,** 1450–1462 (2010).
24. Qiu, X. & Janson, C. A. Structure of apo acyl carrier protein and a proposal to engineer protein crystallization through metal ions. *Acta Crystallogr. D* **60,** 1545–1554 (2004).
25. Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wuthrich, K. TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc. Natl Acad. Sci. USA* **95,** 13585–13590 (1998).
26. Hansen, M. R., Mueller, L. & Pardi, A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nature Struct. Biol.* **5,** 1065–1074 (1998).
27. Markwick, P. R. L. & McCammon, J. A. Studying functional dynamics in bio-molecules using accelerated molecular dynamics. *Phys. Chem. Chem. Phys.* **13,** 20053–20065 (2011).
28. Markwick, P. R. *et al.* Toward a unified representation of protein structural dynamics in solution. *J. Am. Chem. Soc.* **131,** 16968–16975 (2009).
29. Frueh, D. P. *et al.* Dynamic thiolation–thioesterase structure of a non-ribosomal peptide synthetase. *Nature* **454,** 903–906 (2008).
30. Alekseyev, V. Y., Liu, C. W., Cane, D. E., Puglisi, J. D. & Khosla, C. Solution structure and proposed domain–domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci.* **16,** 2093–2107 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements M.D.B. and S.-C.T. are supported by GM100305 and GM095970. We thank J. J. LaClair for figure editing. We thank X. Huang for assistance with NMR facilities and experimental setup. Portions of this research were carried out at the Stanford Synchrotron Radiation Lightsource (SSRL), a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The Advanced Light Source is supported by the Office of Basic Energy Sciences of the US Department of Energy under contract no. DE-AC02-05CH11231. J.A.M. is supported by NSF, NIH and HHMI. Portions of the research are supported by the Advanced Light Source, supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-AC02-05CH11231.

Author Contributions C.N., assisted by G.C.-F., D.R.J. and J.B., determined the AcpP–FabA X-ray crystal structures. R.W.H., D.J.L. and B.O. conducted the protein NMR experiments under the supervision of S.J.O. F.I. and K.F. prepared the crosslinking probe. P.R.L.M. conducted molecular dynamics simulations under the supervision of J.A.M. C.N., G.C.-F., R.W.H. and D.J.L. analysed data and contributed to writing of the paper. S.-C.T. and M.D.B. directed the research, provided funding and wrote the final manuscript.

Author Information The atomic coordinates of AcpP–FabA have been deposited in the Protein Data Bank under accession 4KEH. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.D.B. (mburkart@ucsd.edu) or S.-C.T. (sctsai@uci.edu).

LETTER

doi:10.1038/nature12816

Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection

Lei Sun[1]*, Lindsey N. Young[2]*, Xinzheng Zhang[1]*, Sergei P. Boudko[1]†, Andrei Fokine[1], Erica Zbornik[1], Aaron P. Roznowski[2], Ian J. Molineux[3], Michael G. Rossmann[1] & Bentley A. Fane[2]

Prokaryotic viruses have evolved various mechanisms to transport their genomes across bacterial cell walls[1]. Many bacteriophages use a tail to perform this function, whereas tail-less phages rely on host organelles[2–4]. However, the tail-less, icosahedral, single-stranded DNA ΦX174-like coliphages do not fall into these well-defined infection processes. For these phages, DNA delivery requires a DNA pilot protein[5]. Here we show that the ΦX174 pilot protein H oligomerizes to form a tube whose function is most probably to deliver the DNA genome across the host's periplasmic space to the cytoplasm. The 2.4 Å resolution crystal structure of the *in vitro* assembled H protein's central domain consists of a 170 Å-long α-helical barrel. The tube is constructed of ten α-helices with their amino termini arrayed in a right-handed super-helical coiled-coil and their carboxy termini arrayed in a left-handed super-helical coiled-coil. Genetic and biochemical studies demonstrate that the tube is essential for infectivity but does not affect *in vivo* virus assembly. Cryo-electron tomograms show that tubes span the periplasmic space and are present while the genome is being delivered into the host cell's cytoplasm. Both ends of the H protein contain transmembrane domains, which anchor the assembled tubes into the inner and outer cell membranes. The central channel of the H-protein tube is lined with amide and guanidinium side chains. This may be a general property of viral DNA conduits and is likely to be critical for efficient genome translocation into the host.

Bacteriophage ΦX174 is a small icosahedral microvirus with a circular single-stranded (ss)DNA genome[6]. Electron microscopic and crystallographic studies have shown that the virions have spikes on each pentameric vertex[7]. The capsid, consisting of 60 F proteins, has an external diameter of about 260 Å. Each of the 12 spikes consists of five G proteins, which protrude 32 Å above the F-protein shell. The capsid also contains 60 copies of the DNA-binding protein J and 10–12 copies of the DNA pilot protein H[8]. Although a small portion of each H protein may be in the hydrophilic channel formed by the G proteins[7], the structure and location of the H proteins remained unknown as all structure determinations assumed icosahedral symmetry. The mature ΦX174 virion is devoid of a visible external tail.

Capsids of tailed double-stranded (ds)DNA bacteriophages have a special vertex occupied by a dodecameric portal protein. Myoviruses, such as phage T4, have a contractile sheath surrounding the tail tube that contracts after the virus detects a host cell surface, thereby causing the specialized tip of the inner tail tube to puncture the outer membrane. The tail-associated lysozyme then digests the peptidoglycan layer before a tube component punctures the inner membrane. The genome then passes through the tail tube into the cytoplasm[9]. Podoviruses, such as T7, ε15 and P22, eject internal head proteins through their portals that may also digest the peptidoglycan layer in the cell wall. Afterwards, they either assemble translocation tubes or lengthen their existing short tail tubes to stretch across the periplasmic space[10,11]. By contrast, filamentous and tail-less icosahedral bacteriophages, with the notable exception of the microviruses, evolved to use host-cell-encoded channels[3,4].

The ΦX174 DNA pilot protein H guides the DNA through the cell wall in the penetration process[5]. The genome-associated H proteins are transported to the cytoplasmic membrane, whereas the coat and spike proteins remain outside the cell[12]. Because ΦX174 does not have a tail, Jazwinski and colleagues speculated that the H proteins might form a DNA translocating channel through the cell wall[13]. The H protein contains a predicted N-terminal transmembrane helix, rich in the (G/A)XXX(G/A) sequences known to promote membrane protein oligomerization[14,15]. The protein's central core has a potential coiled-coil structure[16].

In attempts to crystallize the H protein (328 amino acids) it was necessary to remove some of the N- and C-terminal residues. The first successful structure solution was of a fragment stretching from residues 143 to 211 (143H211), which was determined by single-wavelength anomalous diffraction phasing. The structure consists of ten α-helices organized into an α-helical barrel. Using this as a model, a second longer structure (143H282) was determined by molecular replacement (Supplementary Table 1). Like the 143H211 structure, the structure of 143H282 consists of ten nearly identical α-helical monomers that form a cylindrical α-helical barrel. The cylinder is 170 Å long and has an external diameter of approximately 48 Å. The α-helices run parallel along the entire cylinder. Each monomer is kinked at residues Y193–A194–Q195, which divides the barrel into N- and C-terminal domains (Fig. 1a). Domain A (residues 144–194), which is very slightly conical, has a minimum internal diameter of approximately 22 Å, whereas the internal diameter of domain B (residues 195–272) is approximately 24 Å (Fig. 1b). The inner surface of the H tube has mostly negative electrostatic potential (Fig. 1c). To our knowledge, this is the first description of a decameric coiled-coil structure[17].

Most coiled-coil structures are dimers or trimers. The interaction between helices is determined by the insertions of side chains, or knobs, from one helix into the holes formed between side chains of the surrounding helices[18]. In a structure with more than four identical helices, each helix interacts only with its two nearest neighbours. As the number of participating helices increases, the diameter of the central cylindrical cavity expands. This results in an empty tube with a water-accessible cavity, as opposed to the interacting hydrophobic surfaces observed in typical coiled-coil structures[19]. In the H-protein tubes, the interaction between neighbouring helices is dominated by hydrogen bonds (Supplementary Table 2).

The number of residues in one turn of an α-helix is non-integral and is expressed as the number of residues (n) in a given number of complete turns (t). Thus, the number of residues in a single helical turn is n/t. A conventional straight α-helix has 3.6 residues per turn[20]. However, if $n/t > 3.6$ the helix forms a right-handed superhelix, and if $n/t < 3.6$ the superhelix is left-handed[19]. In the present structure, the

[1]Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA. [2]School of Plant Sciences and the BIO5 Institute, University of Arizona, Tucson, Arizona 85721, USA. [3]Molecular Genetics and Microbiology, Institute for Cell and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA. †Present address: The Research Department, Shriner's Hospital for Children, Portland, Oregon 97239, USA.
*These authors contributed equally to this work.

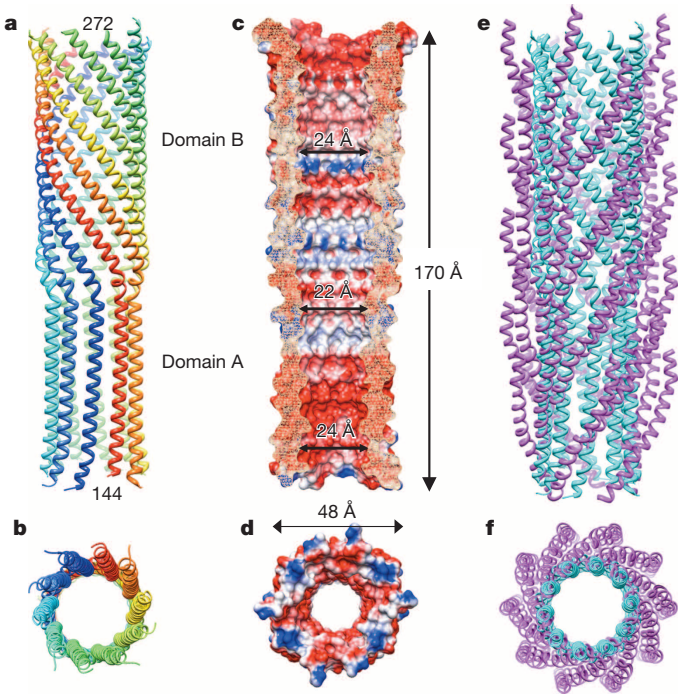


Figure 1 | The structure of the H-protein coiled-coil tube. **a**, Ribbon representation of the structure. The monomers are individually coloured. **b**, Top view of the structure with the N terminus closest to the viewer. **c**, The inner surface of the H-protein tube is coloured according to the electrostatic potential. Blue and red colours correspond to positive and negative potential of 5 kT/e^{-}, respectively. **d**, The top view of the tube surface (the orientation of the tube is the same as in **b**). **e**, **f**, Superposition of the H-protein tube (cyan) and the bacteriophage fd capsid (purple), showing that the H protein has a similar inner diameter to fd (Protein Data Bank accession number 2HI5)[27].

α-helices have 11 residues in three turns (11/3) in domain A and seven residues in two turns (7/2) in domain B. The inclination of the α-helices to the central cylinder axis in domains A and B is 5° right-handed and 36° left-handed, respectively (Fig. 1a).

In domain A, the 11 residues per structural repeat are identified by the letters *a* to *k* (Fig. 2b). The residue that has its side chain pointing towards the centre was arbitrarily designated residue *a*. Within the structural repeat, the side chain of residue *h* would be similarly oriented. Seven of the eight residues at positions *a* or *h* are glutamines. In domain B there are seven residues per turn identified by the letters *a* to *g* (Fig. 2a). Only one amino-acid side chain within the heptad repeat points exactly towards the tube's centre; again, this is designated as position *a*. However, the long side chains of residues *d* and *e* may also assume inward orientations (Fig. 2c). Most of the inward-facing residues are glutamines. A few of them are asparagine and arginine residues. The occurrence of glutamine, asparagine and arginine residues that point towards the centre of the cylinder is also conserved in the homologous pilot protein of other microviruses (for example, coliphages α3 and G4) as well as the distantly related gokushoviruses (for example, *Bdellovibrio* phage ΦMH2K and *Chlamydia* phage Chp2). These long side chains point to the N termini, which may suggest that the DNA is transported from the C to N termini. Like an outside-in harpoon, the side chains, which point to the N terminus of an α-helix, may prevent the DNA from going backwards.

During assembly the external scaffolding protein D organizes twelve 12S* particles into the procapsid. The 12S* particle consists of five copies each of proteins F, G and internal scaffolding B, but only one copy of protein H[21]. Thus, if H-protein tubes are a component of the viral life cycle, some initial H-protein assembly most probably occurs after procapsid formation but before genome packaging. If this is correct, mutations that inhibit the requisite tube forming helix–helix interactions should not affect procapsid assembly. However, if tube formation is essential for penetration of the genome into the host's cytoplasm, the resulting particles should lack infectivity. Using the atomic structure as a guide (Fig. 2), residues that make monomer-to-monomer contacts were mutated. Mutant 1 (N161S + N187A) contains serine and alanine substitutions. Mutant 2 (N161S + N187A + Q224A) contains an extra alanine substitution (Supplementary Table 2 and Fig. 2).

To determine whether the mutant H proteins affect particle formation and/or infectivity, particles were purified from cells infected with mutant or wild-type viruses. The assembled particles were purified by rate zonal sedimentation. Wild-type and mutant virions sedimented at the same rate (Extended Data Fig. 1a). The mutant H proteins were incorporated into virions at wild-type levels (Supplementary Table 3). Mutant particles did not have an altered DNA content, as determined by $D_{260\,nm}/D_{280\,nm}$ ratios (Supplementary Table 3). However, the mutant particles showed approximately three orders of magnitude lower specific infectivity (plaque-forming units/$D_{280\,nm}$) than the wild-type virus (Supplementary Table 3).

The ability of the mutant H proteins to form tubes was investigated *in vitro*. Cloning, gene expression and protein purification protocols were identical to those used for the wild-type H protein. However, size-exclusion chromatography showed that the mutant proteins migrated as lower-order oligomers (Extended Data Fig. 1b). The mutant H proteins were also probed by limited trypsin proteolysis. Whereas digestion of the wild-type 143H282 fragment generated the stable 143H221 fragment, the mutant proteins were digested into small pieces (Extended Data Fig. 1c). This indicated that some lysine and arginine residues, which are buried in the wild-type 143H221 fragment, are accessible in the mutant proteins, implying the failure of the mutant proteins to form tubes. Thus, the biophysical characterization of the mutant particles *in vivo* and mutant H proteins *in vitro* suggest that H-protein oligomerization is critical for infectivity but not for particle morphogenesis.

The H proteins self-assemble into a tube with dimensions suitable for DNA translocation and long enough to span the periplasmic space (Fig. 3) or a Bayer's patch, a membrane adhesion site where infecting ΦX174 particles have been suggested to congregate[22]. Secondary structural prediction indicated a transmembrane helix at the N terminus[14]. Both N- and C-terminal regions are rich in alanine, glycine and serine residues, which have high occurrence in transmembrane helices (Supplementary Table 4). These regions may anchor the H tube in the inner and outer membranes. Tubular structures are also used for transporting the genome to the host's cytoplasm by tailed bacteriophages such as T4 (ref. 9) and T7 (ref. 10), although for T4 the tube is attached to the head before infection whereas for T7 it is extruded from the head at the time of infection. For ΦX174 it is unclear whether the H monomers begin to assemble as the genome is packaged or whether the tubes are assembled from H monomers at the time of infection (Extended Data Fig. 2). In the same way that tailed dsDNA bacteriophages use a virally encoded ATPase to package genomes into phage heads[23], ΦX174 uses a host-cell-derived enzyme[24] to package DNA into the phage procapsids. Presumably some of this packaging energy can subsequently be used to deliver the genome to the next host cell.

As the ΦX174 genome is a circular ssDNA molecule, the H tubes must accommodate two oppositely oriented DNA strands. The packaged genome contains mostly unpaired bases[25]. The minimum internal diameter of the H tube is approximately 22 Å, which can easily accommodate two unpaired ssDNA strands with intercalated bases[26]. Similarly, the circular ssDNA genomes of filamentous bacteriophages (for example, M13 and fd) are packaged into cylindrical fivefold symmetric capsids (Fig. 1e, f) with comparable inner dimensions[3,27].

An electron microscopic tomographic investigation was used to detect the suspected H tubes crossing the periplasmic space of *Escherichia coli* C during ΦX174 infection. However, as normal size *E. coli* cells are too large to allow electrons to cross while forming an image of such a cell, it was necessary to use 'mini' *E. coli* K12 cells[11]. In turn, that required the use of the homologous and closely related microvirus ST-1 (Extended

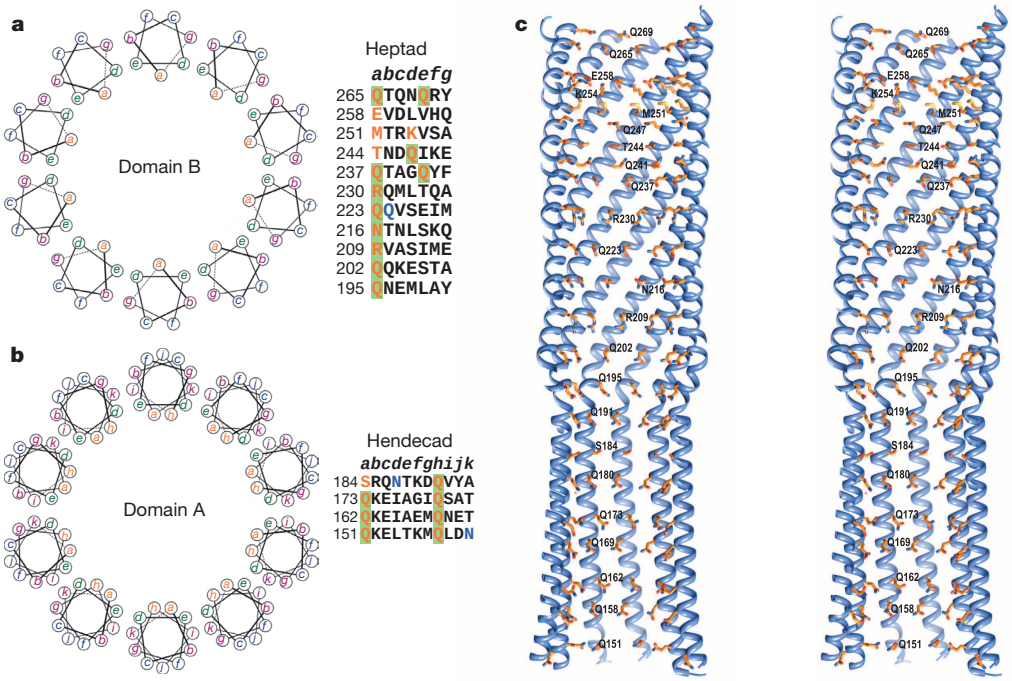


Figure 2 | Helical wheel representations summarizing the inter-helical contacts in different parts of the helical barrel. **a**, The 7/2 heptad repeats of domain B. **b**, The 11/3 hendecad repeats of domain A. The amino-acid sequences arranged in 11- and 7-residue repeats are given on the right. The residues that were mutated to disrupt formation of the H tube are coloured blue. Residues facing the interior are coloured orange. The inward-facing residues are highlighted in green. **c**, Stereo view of the residues lining the tube centre.

Data Fig. 3) as ΦX174 does not infect K12 cells[28]. The ST-1 particles were seen to attach to the surface of the cell after incubating for 6 min or more (Fig. 3a–c). About 20% of the attached particles had a tube emerging from the capsid and, in most cases, crossing from the outer to the inner membrane of the cell (Fig. 3e, f). However, empty attached virions had lost the tube after the virus had delivered its genome to the host cell (Fig. 3g, h).

To our knowledge, the H-protein structure represents the first virally encoded, cell-wall-spanning, DNA-translocating conduit determined at atomic resolution. Many bacteriophage protein assemblies that function to conduct DNA have similar properties. For example, a domain of the P22 portal protein forms a 140-Å-long α-helical barrel, which shares sequence similarity with the ΦX174 H-protein coiled-coil domain (Extended Data Fig. 4)[29]. Although this domain is located within the capsid, dsDNA passes through it during DNA packaging and penetration of the cell wall. Subsequently, the P22 DNA pilot protein, gp16, which contains a predicted coiled-coil domain, transports the DNA across the cell wall[30]. Similarly, the gp14 and gp15 proteins of bacteriophage

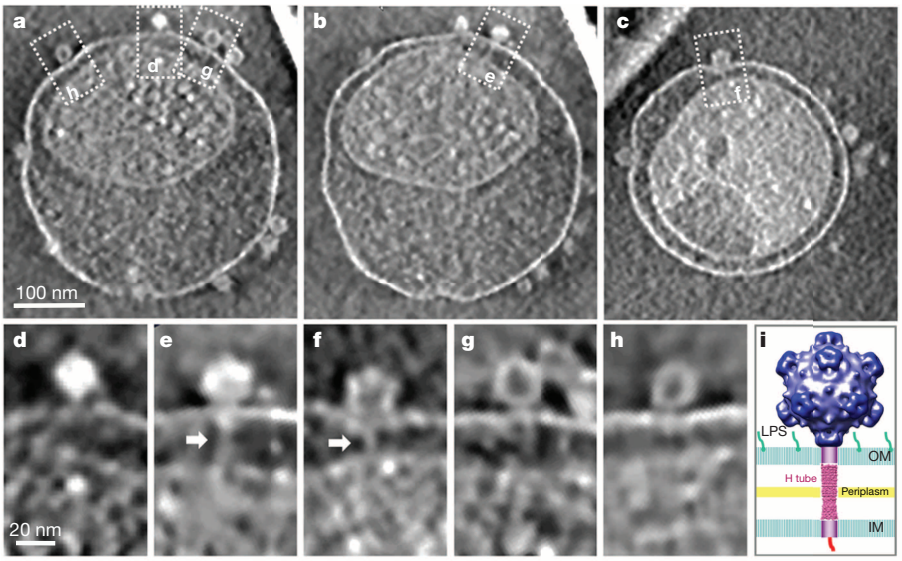


Figure 3 | Cryo-electron micrographic tomogram of the ΦX174-like phage ST-1 infecting *E. coli* mini cells. **a**–**c**, Slices of tomograms showing three states of the infection process. **d**–**h**, Enlarged images taken from **a**–**c**. **d**, The virus has attached to the outer membrane (OM). One of the pentameric spikes of an icosahedral particle has recognized a lipopolysaccharide (LPS) molecule in the outer membrane of the *E. coli* cell wall. **e**, **f**, After attachment, the virus extrudes a tube for DNA penetration. A tube can be seen (white arrow) crossing the periplasmic space, lodged in the outer and inner membrane (IM). **g**, **h**, After DNA has been injected into the cell, the extended tail starts to disassemble. **i**, Schematic model of ΦX174 infection.

T7 that constitute the trans-cell-wall extended tail tube[11] contain strongly predicted coiled-coil domains. The amino-acid sequence of gp14 is similar enough to that of the H protein to allow a reasonable sequence alignment (Extended Data Fig. 5). The spacing of the amide containing side chains, which line the H tube's inner cavity, is similar.

Like other phages, ΦX174 seems to have a 'tail' that is required for infecting a host, but it protrudes from the virion only at the time of infection. Just-in-time proteinaceous or lipidic tail formation also occurs when the T7 (ref. 11) and PRD1 (ref. 4) phages infect cells, respectively.

METHODS SUMMARY

The H-protein segment, corresponding to residues 143–282 (143H282), was cloned and expressed in *E. coli*. The purified protein was digested by trypsin and a smaller stable fragment, 143H221, was further purified. The Se-met 143H221 fragment was crystallized and the structure was determined by single-wavelength anomalous dispersion. The structure of the longer crystallized fragment 143H282 was then determined by molecular replacement, using the structure of 143H221. Missense mutations in the ΦX174 H gene were generated by oligonucleotide-mediated mutagenesis. Mutated virus was extracted from cells that had been infected with mutated ΦX174 DNA. The 143H282 protein mutants were cloned, expressed and purified in the same way as the wild-type 143H282 protein. Tomographic images of the homologous virus ST-1 infecting mini *E. coli* cells were taken on an FEI Titan Krios electron microscope.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 May; accepted 28 October 2013.
Published online 15 December 2013.

1. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2,** a000414 (2010).
2. Molineux, I. J. & Panja, D. Popping the cork: mechanisms of phage genome ejection. *Nature Rev. Microbiol.* **11,** 194–204 (2013).
3. Russel, M. & Model, P. in *The Bacteriophages* (ed. Calendar, R.) 146–160 (Oxford Univ. Press, 2006).
4. Peralta, B. *et al.* Mechanism of membranous tunnelling nanotube formation in viral genome delivery. *PLoS Biol.* **11,** e1001667 (2013).
5. Jazwinski, S. M., Lindberg, A. A. & Kornberg, A. The gene H spike protein of bacteriophages ΦX174 and S13. I. Functions in phage-receptor recognition and in transfection. *Virology* **66,** 283–293 (1975).
6. Sinsheimer, R. L. A single-stranded DNA from bacteriophage ΦX174. *Brookhaven Symp. Biol.*, 12, 27–34 (1959).
7. McKenna, R. *et al.* Atomic structure of single-stranded DNA bacteriophage ΦX174 and its functional implications. *Nature* **355,** 137–143 (1992).
8. Burgess, A. B. Studies on the proteins of ΦX174. II. The protein composition of the ΦX coat. *Proc. Natl Acad. Sci. USA* **64,** 613–617 (1969).
9. Leiman, P. G., Chipman, P. R., Kostyuchenko, V. A., Mesyanzhinov, V. V. & Rossmann, M. G. Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell* **118,** 419–429 (2004).
10. Chang, J. T. *et al.* Visualizing the structural changes of bacteriophage epsilon15 and its *Salmonella* host during infection. *J. Mol. Biol.* **402,** 731–740 (2010).
11. Hu, B., Margolin, W., Molineux, I. J. & Liu, J. The bacteriophage T7 virion undergoes extensive structural remodeling during infection. *Science* **339,** 576–579 (2013).
12. Incardona, N. L. & Selvidge, L. Mechanism of adsorption and eclipse of bacteriophage ΦX174. II. Attachment and eclipse with isolated *Escherichia coli* cell wall lipopolysaccharide. *J. Virol.* **11,** 775–782 (1973).
13. Jazwinski, S. M., Marco, R. & Kornberg, A. The gene H spike protein of bacteriophages ΦX174 and S13. II. Relation to synthesis of the parenteral replicative form. *Virology* **66,** 294–305 (1975).
14. Tusnady, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17,** 849–850 (2001).
15. Russ, W. P. & Engelman, D. M. The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.* **296,** 911–919 (2000).
16. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled-coils from protein sequences. *Science* **252,** 1162–1164 (1991).
17. Woolfson, D. N., Bartlett, G. J., Bruning, M. & Thomson, A. R. New currency for old rope: from coiled-coil assemblies to α-helical barrels. *Curr. Opin. Struct. Biol.* **22,** 432–441 (2012).
18. Crick, F. H. C. The packing of α-helices: simple coiled-coils. *Acta Crystallogr.* **6,** 689–697 (1953).
19. Gruber, M. & Lupas, A. N. Historical review: another 50th anniversary – new periodicities in coiled coils. *Trends Biochem. Sci.* **28,** 679–685 (2003).
20. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl Acad. Sci. USA* **37,** 205–211 (1951).
21. Cherwa, J. E., Jr, Organtini, L. J., Ashley, R. E., Hafenstein, S. L. & Fane, B. A. *In vitro* assembly of the ΦX174 procapsid from external scaffolding protein oligomers and early pentameric assembly intermediates. *J. Mol. Biol.* **412,** 387–396 (2011).
22. Bayer, M. E. & Starkey, T. W. The adsorption of bacteriophage ΦX174 and its interaction with *Escherichia coli*; a kinetic and morphological study. *Virology* **49,** 236–256 (1972).
23. Sun, S. *et al.* The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell* **135,** 1251–1262 (2008).
24. Aoyama, A., Hamatake, R. K. & Hayashi, M. *In vitro* synthesis of bacteriophage ΦX174 by purified components. *Proc. Natl Acad. Sci. USA* **80,** 4195–4199 (1983).
25. Benevides, J. M., Stow, P. L., Ilag, L. L., Incardona, N. L. & Thomas, G. J., Jr. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage ΦX174 determined by Raman spectroscopy: a model for ΦX174 DNA packaging. *Biochemistry* **30,** 4855–4863 (1991).
26. Shepard, W., Cruse, W. B., Fourme, R., de la Fortelle, E. & Prange, T. A zipper-like duplex in DNA: the crystal structure of d(GCGAAAGCT) at 2.1 Å resolution. *Structure* **6,** 849–861 (1998).
27. Wang, Y. A. *et al.* The structure of a filamentous bacteriophage. *J. Mol. Biol.* **361,** 209–215 (2006).
28. Bowes, J. M. & Dowell, C. E. Purification and some properties of bacteriophage ST-1. *J. Virol.* **13,** 53–61 (1974).
29. Olia, A. S., Prevelige, P. E., Jr, Johnson, J. E. & Cingolani, G. Three-dimensional structure of a viral genome-delivery portal vertex. *Nature Struct. Mol. Biol.* **18,** 597–603 (2011).
30. Perez, G. L., Huynh, B., Slater, M. & Maloy, S. Transport of phage P22 DNA across the cytoplasmic membrane. *J. Bacteriol.* **191,** 135–140 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Kelly for help in preparing the manuscript. Use of the Advanced Photon Source (Sector 23) was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract number DEAC02-06CH11357. This research was supported by National Science Foundation grants MCB-0948399 (to B.A.F.) and MCB-0443899 (to M.G.R.) and US Department of Agriculture Hatch funds to the University of Arizona (to B.A.F.).

Author Contributions B.A.F. and M.G.R. developed the concept. L.S., L.N.Y. and X.Z. designed the experiments. L.S. and S.P.D. worked on the cloning, protein purification and crystallization of the H protein. L.S. and A.F. worked on the structure determination and analysis. L.S., L.N.Y. and B.A.F. characterized the mutant data. L.S., X.Z. and B.A.F. produced the tomographic results. B.A.F., I.J.M., E.Z. and A.P.R. contributed effort to protein, virus and cell purification. L.S., M.G.R. and B.A.F. wrote the paper.

Author Information The atomic coordinates and structure factors of ΦX174 H protein have been deposited in the Protein Data Bank under accession numbers 4JPN and 4JPP. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G.R (mr@purdue.edu) or B.A.F. (bfane@email.arizona.edu).

CORRIGENDUM

doi:10.1038/nature12812

Corrigendum: A Jurassic mammaliaform and the earliest mammalian evolutionary adaptations

Chang-Fu Zhou, Shaoyuan Wu, Thomas Martin & Zhe-Xi Luo

Nature **500,** 163–167 (2013); doi:10.1038/nature12429

In this Article, the 'S1-3' label was misplaced in Fig. 1c, and the legend to Fig. 2c should state 'Labial view' rather than 'Lingual view'. In addition, in Supplementary Figs 2d and 3c, the yellow and red labels for upper teeth were the wrong way round, and the legend to Supplementary Fig. 2 should state 'labial' rather than 'lingual' view, and from the 'right side of skull in labial view' rather than 'left side of skull in lingual view'. These changes have been corrected in the HTML and PDF versions of the original paper and its Supplementary Information.

TECHNOLOGY FEATURE

LIFE IN THE DANGER ZONE

Instruments for studying microbes under biological containment cannot be readily removed from labs for servicing. A US facility is finding ways around that problem.

NIH/NIAID

A laboratory in Frederick, Maryland, will soon bring a high degree of automation and imaging capability to research under biosafety-level-4 conditions.

BY VIVIEN MARX

Some bacteria, viruses and toxins are deadly — natural threats to humans and the environment as well as potential bioweapons. To counter such hazards, laboratories that study these pathogens and substances must do so under high security. One such lab is opening this spring in Frederick, Maryland. It will be designated as biosafety-level-4 (BSL-4), the highest level of biological containment.

The lab is part of the Integrated Research Facility (IRF), a complex operated by the US National Institute of Allergy and Infectious Diseases (NIAID) that has been opening in stages since 2008. Although some of the equipment in the 1,020-square-metre lab is standard for a BSL-4 environment, the degree of automation and integration is unprecedented, says Peter Jahrling, director of the IRF.

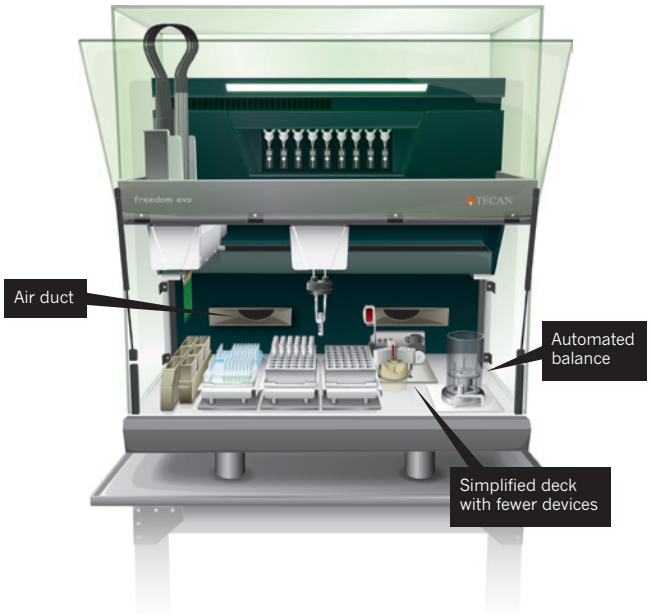
For example, a complete imaging suite and clinical area lets researchers study and treat infected animals without having to remove them from containment. "We've basically built an intensive care unit for animals," says Jahrling. Likewise, the instruments used to screen and study blood and tissue samples have been customized to minimize the need for handling by people. In some cases, almost every step has been automated, with robots doing assays to check virus concentrations or assess immune responses.

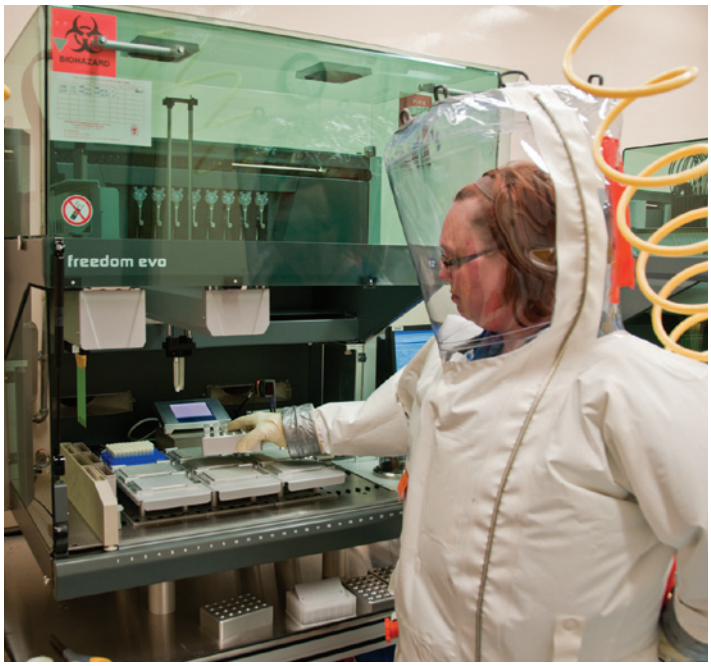
Until the lab receives its final BSL-4 approval, which is expected early this year, its staff is working on slightly less dangerous microbes, such as the viruses that cause cowpox and Middle East respiratory syndrome (MERS). But the researchers are already following BSL-4 procedures — wearing pressurized whole-body suits with air piped in through hoses, and taking disinfecting ▶

SOURCE: JIRO WADA, NIH/NIAID

A CUSTOM JOB

Liquid-handling robots in a biosafety-level-4 lab have to be customized in many ways.





NIH/NIAID

A liquid-handling robot designed for use by scientists in bulky protective gear.

chemical showers in the suits when exiting the lab. "It allows our staff here to gain practice in working with a human pathogen in a BSL-4 environment," says Lisa Hensley, a microbiologist and the IRF's associate research director.

Once the lab has final approval, Hensley and her team will move on to study the Ebola, Hendra and Nipah viruses. "We'll be able to see what's really happening in animals," she says, and to assess potential treatments. The lab's unique ability to image live specimens in the BSL-4 environment will help researchers to cut down on the number of animals killed, and will avoid the need for extensive and numerous post-mortem examinations, which increase the risk to staff.

But installing robots and complex imaging equipment in the lab has called for careful engineering.

CLEAN IMAGING

In 2012, scientists at another NIAID BSL-4 lab, in Hamilton, Montana, showed that the infection that causes MERS in humans could be modelled in macaque monkeys (V. J. Munster *et al. N. Engl. J. Med.* **368,** 1560–1562; 2013). The scientists there could do blood work on these animals and take X-rays, says Hensley, but more detailed imaging data are crucial for identifying lung regions affected by the disease, pinpointing possible locations where inflammation begins and assessing possible treatments. "You can look to see if your countermeasure is working," says Hensley.

That is why the new IRF lab has a hospital-quality imaging suite that includes an X-ray imager, a 3-tesla magnetic resonance imaging machine and scanners that combine computed tomography with positron emission tomography or single-photon emission computed tomography (see 'Hot and cold').

But each piece of equipment has many moving parts and electronic components that need regular servicing, says Jahrling. "There's no way you can put that in a BSL-4 lab and expect it to function for very long," he says. And in a lab with such high biosafety requirements, "the service people aren't coming in", says Hensley. The equipment lives in containment and can leave only if thoroughly decontaminated, yet no instrument can handle a chemical shower. The alternative — treatment with vaporous formaldehyde or hydrogen peroxide — can also cause damage. And even if the instrument were to survive, the process would be time-consuming.

So before the IRF scientists installed imaging equipment in the BSL-4 lab, they collaborated with engineers at Philips Healthcare, which has its US headquarters in Andover, Massachusetts, to completely re-engineer the instruments. "That was a huge research and development effort," says Jahrling.

"The more things you add to a robot, the more things can go wrong."

The main sections of the imaging machines are installed on one side of a wall — the 'cold' or non-contained side. On the 'hot' side is a patient table onto which an anaesthetized animal is placed for imaging. The table is mounted on tracks and can be moved through a secure transparent tube that extends into the cold side of the lab but is sealed to prevent the escape of pathogens. Researchers then slide the table through the tube to the imaging apparatus on the cold side. This design allows service staff to maintain the instrument outside the contained lab.

Four liquid-handling robots have been installed at the IRF, two of which are in the BSL-4 lab. The instruments are used to extract DNA and RNA from tissues and to perform assays that involve titration and staining. The robots cap and uncap tubes, weigh them and dispense various reagents.

In plaque assays, used for calculating the concentration of virus in a sample, virus is added to a multiple-well plate that contains cells and culture medium. After a certain amount of time, visible holes or 'plaques' form where the virus has infected and killed cells. The assay requires multiple steps that include serial dilution, weighing samples and adding buffers.

In neutralization assays, used for assessing immune responses, scientists can see how well antibodies interfere with an infectious agent. When robots tend to these assays, there is less variability than when a number of different staff members perform the task, says Hensley. Less handling of infectious and toxic agents is also safer for scientists, she says.

SIMPLIFIED ROBOTS

Automation brings its own problems. For example, Jahrling says, "there's a concern that the repetitive manipulations that the robot performs might generate an aerosol" — tiny, airborne droplets that can spread infectious agents. To reduce the risk, some instruments were enclosed and tested to ensure that performance was not impaired.

Kenny Ung, an engineer at Tecan, a company based in Männedorf, Switzerland, and his colleagues worked with the IRF team to customize Tecan's Freedom EVO liquid handlers for use in the lab. The robots are enclosed by safety shields.

Tecan customizes instruments for complex workflows in which robotic arms move between carousels, bar-code scanners, incubators, shakers, sealers and other devices. But the machines needed to be simplified and customized for use in this lab. Given the space constraints and difficulty servicing instruments, Ung says, "you don't want one of those types of systems inside BSL-4". But the IRF team wanted some degree of automation in their assays so that many samples can be prepared at once. "A robot doesn't get bored, doesn't get tired," says Ung.

In the IRF's BSL-4 lab, robots are programmed for sample preparation with fewer additional devices and with as little human intervention as possible (see 'A custom job'). A robotic arm will pick up tubes containing tissue or blood samples, move them to an automated balance, uncap and cap them and add buffer as needed. Another robot dispenses liquids for serial dilution of samples, which are then transferred to 96-well plates to be incubated and analysed.

Tecan's engineers customized the liquid handler to deal with this succession of tasks, but kept the robot simple. They added an automated balance and a bar-code reader to scan and keep track of the sample tubes, but decided not to include a sample-mixing sonicator — instead, the IRF scientists use an external mixer. "The more things you add to a robot, the more things can go wrong," Ung says — which is true especially in a BSL-4 environment, where repair needs must be kept to a minimum.

Ung knew that scientists would be operating the instrument while wearing protective suits, so he decided to simulate their situation. "I wanted to feel what it's like to be the user inside the lab," he says. The IRF team sent him a safety suit to wear while he worked on the machine.

One of Ung's tasks was to ensure that the equipment had no sharp edges that could puncture the suit. In addition to installing safety shields on the instrument's sides and tops, he filed and sanded the edges on the instrument and on its aluminium stand.

"You just get used to disconnecting your air and walking ten feet down to the next air line and reconnecting."
Peter Jahrling

He also elevated the instrument slightly and added a drip pan so that scientists could capture and clean spills easily. He and his colleagues added air ducts, adjusting the air flow to be strong enough to remove potentially pathogen-laden aerosols but not so strong as to disrupt liquid-dispensing or weighing. Increased air flow can prevent a balance from stabilizing, which would lead to flawed readings.

Before installing the machine, the IRF team visited Tecan's labs to verify that it worked as specified. The instrument was then moved to the BSL-4 lab, where it was reassembled and tested. Tecan worked out a maintenance plan for the instrument. "I'm not able to go in there, so they basically need to have skills on their side to remedy the problem," says Ung. Cindy Allan, a biomedical engineer at the IRF, has learned how to take the robot apart and put it back together. She also learned the tricky task of changing the pipette tips, diluters and syringes. Biothreat analysis calls for genetic and protein-based tests, says Amy Altman, who directs biodefence at Luminex, an instrument manufacturer in Austin, Texas. Antibody-based probes are used to detect proteins such as ricin or *Clostridium botulinum* toxin, for example. And some viruses can have very low concentrations in the blood, so detecting them requires sensitive genomic tests.

NIH/NIAID

FAST FLOW

In the IRF's BSL-4 lab, researchers will be using Luminex's FlexMAP 3D, which is based on flow cytometry, a technique in which lasers are used to count cells. In this instrument, specific types of genetic material or protein are attached to beads coloured with multiple fluorescent dyes at different ratios. The beads move in single file past a laser that emits red light to excite the dyes. The emitted wavelength of light identifies the type of bead. A green laser then excites a 'reporter' dye in the bead that determines which protein or nucleic acid it is attached to. "As each bead gets interrogated by the laser, you can think of each bead as its own little test tube," says Altman.

To screen a sample for ten strains of Ebola virus, for example, researchers use an array of beads that each carry snippets of DNA specific to one of the strains. A blood sample from a patient who succumbed to Ebola can therefore be tested to determine which strain caused that person's death.

The platform works quickly and can distinguish between up to 500 genes or proteins

HOT AND COLD

The imaging suite at a US lab in Frederick, Maryland, has a 'hot' and a 'cold' side for biosafety-level-4 research. Hot (pathogens present) Cold (pathogens not present)

SOURCE: FABIAN DE KOK-MERCADO/NIH/NIAID IRF

1 Single-photon emission computed tomography and computed tomography (SPECT-CT)
2 Magnetic resonance imaging (MRI)
3 X-ray imaging
4 Positron emission tomography and computed tomography (PET-CT)

simultaneously. Such 'multiplexing' means that scientists can check for many signatures of a possible infectious agent at once, Altman says; a 'true positive' for a pathogen would be when all signatures of the disease agent are present. Multiplexing platforms are used in biothreat monitoring, which sometimes entails "high-consequence, high-regret decisions", such as whether or not to shut down an airport, says Altman.

To ensure that the hardware could operate in the IRF's BSL-4 lab, Luminex scientists tested whether the instrument could be safely decontaminated with vaporous hydrogen peroxide and paraformaldehyde. Neither substance is usually kind to instruments, but in the case of the Luminex machine, "it made it through fine", Altman says.

BACTERIAL HIGH ALERT

Scientists in the BSL-4 lab will also be using the BacT/ALERT 3D 60, an automated system made by bioMérieux, based in Marcy-l'Étoile, France, that tests samples of body fluid for fungi, mycobacteria or bacteria — such as *Yersinia pestis*, which causes bubonic plague, or the anthrax culprit *Bacillus anthracis*.

To accomplish this task, the instrument has a built-in incubator that rocks sample bottles, swirling their contents to enhance microbial growth. At the bottom of each bottle is a dab of dried silicone that contains a dye and acts as a pH sensor. When the pH in a bottle changes because of microbial respiration, the sensor's colour changes from blue-green to yellow. A detection system illuminates the bottles periodically and alerts lab staff to any change in sensor colour, says Doug Matthews, product manager at bioMérieux.

"Every ten minutes, it takes a reading and will alert you if it detects growth," Matthews says. Speed is important in hospitals, where this instrument is often deployed, as well as in clinical-testing labs and the pharmaceutical and food industries. But in a BSL-4 lab, rapid detection is even more crucial, he says.

After detection through this automated incubator, scientists can then confirm their findings with other steps. A sample will be grown in an agar dish to identify specific pathogens, for example, or to test how well an antimicrobial agent might work. Researchers might also extract genetic material for sequencing.

The BacT/ALERT 3D 60 was not made specifically for BSL-4 labs, but a number of its design features help it to work in that environment, says Matthews. Despite its compact size, it can handle 60 samples. Its touch screen can be operated by heavily gloved hands. The instrument is designed to be hardy — company records of installed instruments show that the machines do not need major service or repair for an average of 1,000 days, he says. The sample bottles are made from nylon sandwiched between two layers of polycarbonate to resist shattering and spreading infectious cultures. "Keeping glass out of a BSL lab is obviously a key safety feature," Matthews says.

JIRO WADA/NIH/NIAID IRF

In a custom computed-tomography scanner, animals can be imaged without contaminating equipment.

Later this year, bioMérieux intends to launch a new model of the instrument, called the Virtuo, in Europe. The Virtuo handles more bottles than the BacT/ALERT 3D 60 and automatically places them in racks after scanning their bar codes. The instrument is now in clinical trials, says Matthews, and the company is slated to file for approval with the US Food and Drug Administration in 2015.

CELLULAR CONNECTION

The IRF's new BSL-4 lab is subdivided into rooms for different tasks, such as sample preparation, cell culture, animal care or imaging, and tests for clinical pathology, molecular biology and virology. But moving between those rooms is not simple, because scientists in protective gear are tethered by an air hose that stretches only so far. To venture farther, they have to detach from their air supply. "If you disconnect your air you could last for ten minutes," says Jahrling. "You just get used to disconnecting your air and walking ten feet down to the next air line and reconnecting."

Given the multitude of tasks researchers do in the IRF lab, they need to confer often. Years of working in 'space suits' have made Hensley and Jahrling fluent in non-verbal communication — by a wave of a gloved hand, for example. For the new lab, they wanted to go one better — in part because they noted a cultural shift among younger scientists, who seemed to want to communicate more when in the BSL-4 lab.

After testing various communication systems, they chose mobile phones and Bluetooth headsets. The phones are external to the suits; the headsets are internal. "You pick up the cell phone on the hot side of the suit," Hensley says.

Communication and collaboration have been a goal for the lab since 2005, when construction of the IRF began as part of a plan to bolster US biodefence research. To enable collaboration, the IRF is located next to the US Army Medical Research Institute of Infectious Diseases (USAMRIID) and labs run by the US Department of Homeland Security. Both Jahrling and Hensley previously worked at the USAMRIID.

Jahrling says that those neighbouring labs are the subject of much rumour and speculation, but that no secret research is done at the IRF. On the contrary, he has extended an open invitation to scientists from other biosafety research facilities to visit, so that they can see for themselves what it is like to battle the most noxious weapons from nature's arsenal in an up-and-running, next-generation BSL-4 lab. ■

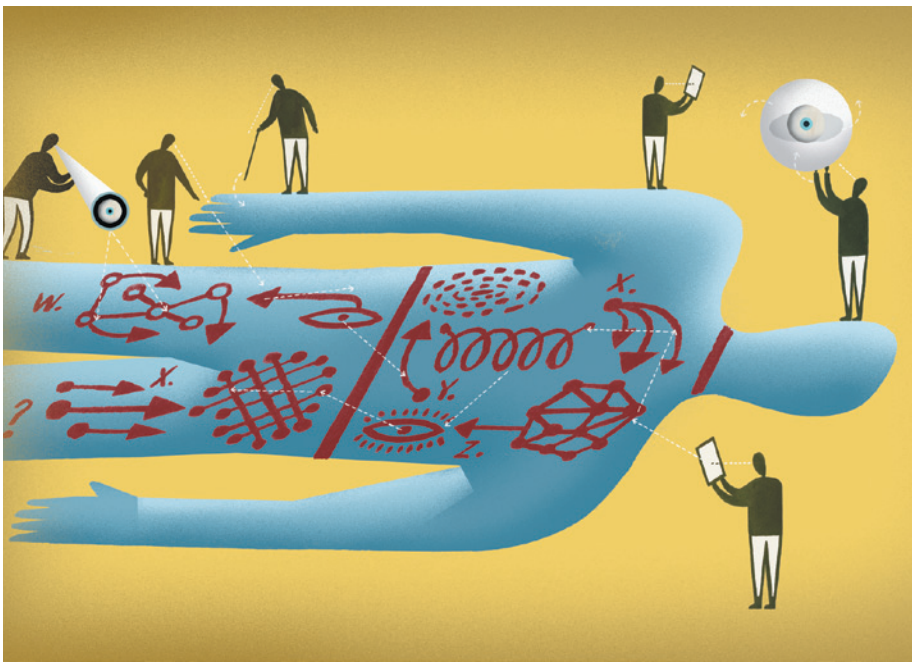
Vivien Marx *is technology editor for* Nature *and* Nature Methods.

GENDER Conference committees with women enlist more female speakers **p.445**

SALARIES Prospective postgraduates ignore future salary outlook **p.445**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

DRAWN IDEAS/IKON IMAGES/CORBIS



DISEASE RESEARCH

Rare insights

Scientists who specialize in uncommon diseases can find a research focus with a purpose.

BY HEIDI LEDFORD

Steven Gray used to spend long hours in the lab for the simple love of science. As a postdoctoral researcher, he was tinkering with a virus in search of ways to shuttle genes into nerve cells for gene therapy. Then, in 2008, his adviser sent him to a meeting held by a non-profit organization called Hannah's Hope Fund, and Gray found a new inspiration.

Hannah's Hope Fund is a charity based in Rexford, New York, that supports research on giant axonal neuropathy (GAN), a fatal nerve disorder. At the meeting, Gray met Hannah Sames, a clumsy four-year-old with tight curls and a sweet smile whose disease had inspired her parents to start the charity. He launched a GAN project after the meeting. "I looked at her and saw my own daughter," says Gray, whose child was then also four. "Now I'm focused on finding a treatment, almost as I would for my own child."

GAN is one of more than 6,000 rare, or 'orphan', diseases that affect humans worldwide. Such diseases typically strike fewer than 1 in 2,000 people, and present unique challenges to researchers and drug developers, who have access to limited numbers of participants for clinical trials and few resources such as animal models.

But for scientists who can overcome such challenges, the rewards can be tremendous. From a practical perspective, an increasing interest from industry and available government funds dedicated to rare diseases have brought new job options. "There are great opportunities for people in academia to interact with pharma and to access government funding," says Daniel Ory, who studies Niemann–Pick type C disease, a genetic neurodegenerative disorder, at Washington University in St. Louis, Missouri.

Rare-disease research also offers rich returns for scientists who thrive on interaction, adds Ory. Because of the small numbers affected and a dearth of information about most rare diseases, scientists must work closely with patients and their families. They also frequently collaborate with patient-advocacy organizations to gather tissue samples, learn more about symptoms and recruit subjects for clinical trials. Personal interaction presents challenges: many rare diseases are fatal, and they often affect children. "It really doesn't get any more raw or emotional," says Nick Leschly, president of biotechnology firm bluebird bio in Cambridge, Massachusetts, which is seeking treatments for several rare diseases.

HIRING OUTLOOK

Thanks mostly to the fervour of patient advocates, research into rare diseases is booming. Philanthropic donations have allowed universities to set up rare-disease centres, such as the University of Pennsylvania's Center for Orphan Disease Research and Therapy in Philadelphia, which has awarded more than US$4.1 million in grants to its researchers since it opened in 2011. In 2012, the US National Institutes of Health (NIH) awarded $3.6 billion for rare-disease research, including supporting dedicated initiatives such as its Therapeutics for Rare and Neglected Diseases programme. The European Commission's Seventh Framework Programme for research funding spent an estimated €530 million (US$720 million) on orphan diseases between 2007 and 2013, according to EURORDIS, an alliance of rare-diseases patient organizations based in Paris.

But it is in industry that the field has really taken off. Pharmaceutical companies that once shied away from developing drugs for small markets have learned from success stories such as Genzyme, an orphan-disease company based in Cambridge, Massachusetts. Genzyme built a booming business, compensating for the small market by charging high prices. (There are programmes to help patients to pay for their drugs, but the pricing remains controversial — see 'Cost conundrum'.) Other companies are now flocking to take advantage of regulatory incentives: in the United States, firms sometimes receive tax credits for clinical trials of orphan-disease drugs, and US and European regulators often streamline the approval of such medicines. One-third of the 39 drugs approved by the US Food and Drug Administration (FDA) in 2012 were for orphan diseases, and the global market for them is expected to grow from $86 billion in 2012 to $112 billion in ▶

2017, according to BCC Research, a market-research firm in Wellesley, Massachusetts.

As a result, orphan-drug development is teeming in small biotechnology firms such as bluebird bio, which almost doubled its staff by adding about 40 new employees this year, most of them scientists. Large pharmaceutical companies, including London-based GlaxoSmithKline (GSK), have developed specialized units that focus on rare diseases. Hans Schikan, chief executive of Prosensa, a rare-diseases company in Leiden, the Netherlands, was surprised when GSK approached him in 2009 about investing in his company's research, because he had not realized that GSK had any interest in rare diseases. The resulting partnership brought Prosensa $25 million, with added payments as the company hit milestones in its Duchenne muscular dystrophy projects. Since then, says Schikan, the field has blossomed. Other large pharmaceutical companies such as Novartis in Basel, Switzerland, have increased their focus on rare diseases, creating job opportunities. "It's still a fairly young space," says Schikan. "But it's growing."

MULTIPLE NICHES

There are many different kinds of orphan disease, and interested researchers can come from a variety of disciplines. Veterans in the field recommend that prospective researchers brush up on statistics because small sample sizes can require sophisticated analyses. Experience in the latest genome-sequencing techniques is a boon: the technology has unearthed the genetic basis for many rare diseases in recent years. "We're seeing many of the tools and technologies developed for common diseases now being applied for the first time in rare diseases," says Robert Steiner, executive director of the Marshfield Clinic Research Foundation in Wisconsin.

And orphan diseases provide an opportunity to carve out a niche and make contributions with little fear of being scooped, Steiner adds. "There is often less competition," he says. "And I've felt as if everything I was doing had the potential to add significant knowledge to an area where there was a real gap."

But carving out a niche can be isolating, as Josh Sommer learned first-hand. In 2006, as an undergraduate at Duke University in Durham, North Carolina, Sommer was diagnosed with chordoma, a rare spinal cancer. He later decided to work in a lab that studied the disease, and Duke, as it happened, was the country's only university with a federally funded chordoma lab. But it did not take long for Sommer to learn about the frustrations of rare-disease research. "We didn't have access to tissue samples, cell lines or mouse models," he says. "And we didn't have other labs to reach out to or collaborate with. It was lonely."

Sommer left the university after his third year to co-found the Chordoma Foundation, also based in Durham, where he remains executive director. One of his first initiatives was an annual workshop to bring together chordoma researchers and other scientists whose work may have a bearing on the disease. The foundation also sponsors $10,000 prizes for scientists who develop useful preclinical models. The cash is not much, says Sommer, but it has incentivized several laboratory technicians to develop chordoma cell lines and mouse models. So far, the foundation has distributed three cell lines to 60 labs and companies, and Sommer says that he stays in contact with more than 150 researchers whose work may be relevant to the disease.

Few preclinical models are available to rare-disease researchers, but this can be a chance to create high-quality models in one's own lab, says Steiner. His biggest worry is making the case for funding. Although the NIH supports a range of orphan-disease research, Steiner says that he and his colleagues still occasionally receive grant reviews questioning the importance of their work. "Some study sections for federal funding agencies are still focused on

AT WHAT PRICE?

Cost conundrum

One way in which companies have made a profit from orphan drugs is by charging high prices — a trend that can frustrate researchers intent on helping patients.

For example, the cystic fibrosis drug Kalydeco (ivacaftor), made by Vertex Pharmaceuticals of Cambridge, Massachusetts, can cost individuals a staggering US$373,000 per year. Drug companies say that they need to charge such high prices to recoup the cost of developing drugs for a tiny number of people. People with health insurance may not have to shoulder much of this cost, and many companies, including Vertex, have global financial-aid programmes to help those without insurance.

But physicians and researchers are voicing concerns about whether these programmes are sufficient for ensuring access, and question how sustainable the high prices are. Charging so much for the drugs disregards the contribution of academic researchers and clinical-trial participants to drug development, argues Carlos Milla, a paediatrician who studies rare diseases at Stanford University in Palo Alto, California. "It's more than just the company and the investors," he says. "There was a whole community that contributed to this effort." H.L.

STEVEN GRAY

Steven Gray (far right) was inspired to launch a project on giant axonal neuropathy, a rare disease.

the idea that the significance of a project is directly proportional to the number of patients affected," he says.

SECURING FUNDS

One method for boosting funding opportunities is to look for ways in which a rare disease overlaps with a more common one. Heather Bean, a chemist at Dartmouth College in Hanover, New Hampshire, is using a two-year postdoctoral fellowship from the Cystic Fibrosis Foundation in Bethesda, Maryland, to support her studies of bacterial lung infections associated with cystic fibrosis. But she hopes eventually to expand her funding opportunities by exploring overlaps between cystic fibrosis and a more common ailment, chronic obstructive pulmonary disease. "I'm still committed to looking at cystic fibrosis," she says. "But drawing those links to a bigger, more fundable disease is handy."

Some labs thrive by pulling in grants and fellowships from multiple foundations. Claudio Hetz, co-director of the Biomedical Neuroscience Institute at the University of Chile in Santiago, studies protein folding, which goes awry in several rare diseases including Creutzfeldt–Jakob disease, Huntington's disease and amyotrophic lateral sclerosis. Hetz was worried about finding funds when he left his postdoc in the United States to open his own lab. He applied to a slew of foundations, hoping to get an award from one. He received three grants. "It was the starting point for everything," he says. "It allowed me to build a solid lab really fast."

Hetz rattles off a list of five foundations that he works with, and says that he has forged personal relationships with people at each of them. Some, such as the Michael J. Fox Foundation for Parkinson's Research, based in New York, do more than just hand over money. Hetz says that he contacts a programme officer there when he encounters a technical stumbling block, and the officer works with him to find the right scientist to consult.

Yet there can be drawbacks to foundation grants. The awards are often smaller than government grants — Gray, who now runs his own laboratory at the University of North Carolina in Chapel Hill, says that the largest of his seven foundation grants is still just shy of $250,000 a year. His smallest have values of about $50,000 per year. And the grants often last only a year or two, creating a sense of instability. Furthermore, foundations rarely pay for full operational costs — such as building maintenance and administrative support — that universities typically take out of government grants. For that reason, universities say that they can lose money on the awards, and sometimes force staff to decline them. Researchers who are interested in competing for foundation money would therefore do well to check with their institutions to find out if they can accept the funds (see *Nature* **504,** 343; 2013).

Foundations also expect their grant recipients to remain focused on the goal of helping patients. Gray warns applicants to his lab that this will sometimes mean dropping scientifically interesting experiments if they do not obviously contribute to the project's main mission. "We really make sure that everything we're doing is in the best interest of the people that are funding us," he says.

Gray is comfortable with that compromise. Last year, he applied to the FDA for approval to conduct a GAN clinical trial. He counts several people with GAN and their families among his friends. His voice is strained when discussing the recent death of an adult with the condition whom he met at that original meeting. "It's tough," he says. "You're always trying to work a little harder." ■

Heidi Ledford *reports for* Nature *from Cambridge, Massachusetts.*

GENDER

Female speakers

Having at least one woman on the speaker-recruiting team for a scientific conference boosts the number of female speakers, finds a 7 January study (A. Casadevall and J. Handelsman *mBio* http://doi.org/qsh; 2014). The authors examined 460 symposia with a total of 1,845 speakers at two annual meetings sponsored by Washington DC's American Society for Microbiology in 2011–2013. They focused on 104 all-male convener teams and 112 with at least one woman. When at least one woman was on the team, the proportion of female speakers rose from an average of 25% to 43%. Co-author Arturo Casadevall, a microbiologist at the Albert Einstein College of Medicine in New York, says that early-career female scientists can benefit from volunteering to be speaker recruiters.

SALARIES

Living in the present

Prospective biomedical postgraduate students decide whether to enrol on the basis of current salaries rather than potential future earnings, says a study out on 23 December (M. E. Blume-Kohout and J. W. Clack *PLoS ONE* http://doi.org/qsq; 2013). Data for 1996–2010 showed that postgraduate enrolment for a given year rose by 2.9–3.9% when relative wages for biomedical-science posts rose by 1%. But enrolment in a given year did not correspond to salary changes six years later, around graduation time. Prospective students should consider effects on salary trends, such as dips in agency budgets, says co-author Meg Blume-Kohout, a senior research economist at the New Mexico Consortium in Albuquerque.

FEMALE RESEARCHERS

Ireland lines up grants

The Irish government plans to launch one- and two-year postdoctoral fellowships, worth up to €185,000 (US$252,000) each, to prompt early-career female researchers to stay in or return to the scientific workforce after childcare or other breaks. Science Foundation Ireland (SFI) in Dublin will announce the 20 or so Advance Fellowship grants by June, says Elena Martines, the SFI's scientific programme officer. Currently, 35% of SFI-funded Irish postdocs and 20% of SFI-funded academic researchers are women, says Martines. "This is a very bold programme," she says, noting that few similar initiatives exist.

SECOND CHANCE

A secure future.

BY KEN LIU

The waitress brings me a glass of carrot juice. She notices that I'm wearing gloves even though it's the middle of summer, but she just shrugs and leaves. Too young to recognize me, I guess. Or maybe she's still embarrassed by how I got kicked out of the game.

I sip the juice through a straw. On the jumbo-sized TV hanging over the bar, they're showing the Red Sox–Yankees game.

A new player steps into the batter's box.

"Look at that stance. It's like watching a historical film," the colour analyst says. "Gives you the shivers."

Herman Ruth settles in, his six-foot-two, 180-pound frame filling the super-hi-def view. He's in prime shape. I'm sure the Red Sox training staff watch his diet like hawks.

"What do you think about the court decision?" asks the play-by-play announcer.

"I think they got it right. I don't see how they can stop him from playing. It's not his fault that he's cloned from Babe Ruth, you know? The kid just loves the game. Maybe he'll have a better record than even the original."

"And you think the small fine is fair?"

"Yeah, I think so. Look, the Red Sox didn't force the kid. They cloned him and found a loving family to adopt him. They didn't tell him who he was and left him alone for 18 years. Then they show up and offer him a job playing a game he loves. I just wish they hadn't disturbed the original Babe Ruth's grave."

"What about the family's objections?"

"Like the court said, your family doesn't own you. You gotta give it to the Red Sox for coming up with this trick to make up for that debacle of a trade a century ago. A second chance for the Sox and the Babe!"

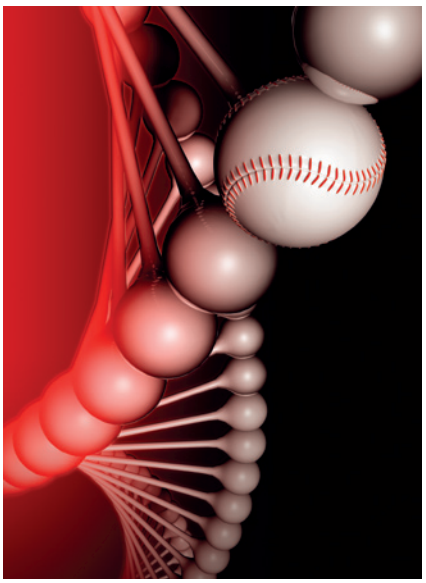
Ruth swings and there's the solid *crack* of wood on leather. He stands there for a fraction of a second, watching the ball sail on its inevitable trajectory out of the ball park. Then his face breaks into that famous grin, and he begins the trot around the bases.

I get up and pocket the straw. I take out some cash with my gloved hands and leave it on the table.

I've been voted the AL MVP three times. All-Star more than a dozen times. I would have had the all-time home-run record. I could have been a great, no, the greatest, player in history.

I won't leave any of my cells behind.

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm



JACEY

"Let me introduce you to Doctor Danzer," my agent Scott says. "She's as good a scientist as I am an agent." I haven't seen Scott in years. In fact, he's been avoiding my calls ever since my disgrace.

I shake the woman's hand. "Pleased to meet you."

"I was a fan," she says. I note the past tense.

"Of all my clients," Scott says in a tone that suggests we're still buddies, "you're the first to bring up the intellectual-property angle. It's brilliant. Once we get the protection in place, we'll sell the cloning rights to the highest bidder."

I harrumph non-committally. He's practically rubbing his hands because I finally found another way to make money for him. A second chance.

"We can't patent your genome, after the Supreme Court said naturally occurring genes can't be patented. We can't copyright it either, as copyright requires a threshold of 'original expression' that's nonfunctional. But I found Doctor Danzer here to help us."

"Directed methylation," she says, and looks at me expectantly.

Scott and I wait politely. But she is apparently finished.

"Doc, you gotta explain more," Scott pleads. "We're not PhDs here."

She sighs. "It's really very simple. As a person develops from a fertilized egg, the DNA gets small bits of hydrocarbons attached to it in a process called methylation, part of gene-expression regulation. It's one of the main ways that the cells in your tissues are different from stem cells."

"How does this help you provide that bit of 'originality' to my genes?" I prod.

"I've invented a technique that will add methyl groups to your DNA in a specific, targeted way. I'll focus on the non-coding segments of your DNA to minimize the possibility of side effects. To make the copyright office really happy, we can use binary code to etch some message into your genes based on the positions of the methyl groups."

"I'll write you a poem."

"That will do. Then we can copyright the whole methylated genome. Since demethylation of somatic cells during cloning tends to be incomplete, if they try to clone you, some of the artificial methylation will survive into the clone."

"Thus infringing my copyright." I'm catching on. "But what if they're willing to pay the damages?"

"Ha!" Scott slaps my back. "This is where it becomes genius. By registering the copyright, we can get statutory damages. How many cells are in a human body, doctor?"

"About a hundred trillion."

"At the minimum statutory damage rate of $750 per copy, that's … many times the national debt. Who can afford to clone you without authorization?"

I can see Scott is already planning on spending his commission.

"You sure about this?" I ask.

"As sure as anything involving epigenetics," Doctor Danzer says. I asked her to stay after Scott left. "I can direct the methylation process in such a way that cloned embryos will not develop properly. You won't be able to sell your copyright at all."

"That'll be perfect," I said. "Thank you, doctor."

"You could have been great," she says.

I shrug.

I don't just want to control my genes for the rest of my life plus 70 years. I want to make sure that there will never be another me in the world, no copies that may excel the original. Call me vain if you want. I may have made my mistakes, but I want to be the only me in the history of the Universe.

No second chances. ■

Ken Liu *is an author and translator of speculative fiction. For more about him and his work, visit http://kenliu.name or follow @kyliu99.*